Structured Sparsity Optimization With Non-Convex Surrogates of $\ell_{2,0}$-Norm: A Unified Algorithmic Framework

Xiaoqin Zhang, *Senior Member, IEEE*, Jingjing Zheng, Di Wang, Guiying Tang, Zhengyuan Zhou, and Zhouchen Lin, *Fellow, IEEE*

Abstract—In this article, we present a general optimization framework that leverages structured sparsity to achieve superior recovery results. The traditional method for solving the structured sparse objectives based on $\ell_{2,0}$-norm is to use the $\ell_{2,1}$-norm as a convex surrogate. However, such an approximation often yields a large performance gap. To tackle this issue, we first provide a framework that allows for a wide range of surrogate functions (including non-convex surrogates), which exhibits better performance in harnessing structured sparsity. Moreover, we develop a fixed point algorithm that solves a key underlying non-convex structured sparse recovery optimization problem to global optimality with a guaranteed super-linear convergence rate. Building on this, we consider three specific applications, i.e., outlier pursuit, supervised feature selection, and structured dictionary learning, which can benefit from the proposed structured sparsity optimization framework. In each application, how the optimization problem can be formulated and thus be relaxed under a generic surrogate function is explained in detail. We conduct extensive experiments on both synthetic and real-world data and demonstrate the effectiveness and efficiency of the proposed framework.

Index Terms—Structured sparsity, non-convex surrogate, fixed-point algorithm

1 Introduction

The massive explosion of available high-dimensional data has become the modern-day norm for a large number of scientific and engineering disciplines and presents a daunting challenge for both computation and learning. Rising to this challenge, sparse recovery techniques have provided a mature framework that exploits the blessing of dimensionality: natural signals are often sparse or compressible in the sense that they have low-dimensional representations when expressed on a proper basis, even though the ambient dimension is often extremely high. Related developments in sparse recovery have thus provided state-of-the-art results in image processing [1], [2], signal processing[3], [4], [5] and machine learning [6], [7], [8].

There are two prominent sparse structures that have been extensively explored in this area: 1) vector sparsity [6], [9], [10], and 2) matric low-rankness [11], [12], [13]. Vector-sparsity-based approaches assume that each observation can be sparsely represented by an over-complete dictionary, and usually use the ℓ_1-norm (i.e., $\|\mathbf{z}\|_1 = \sum_i |z_i|$) as the convex relaxation of the ℓ_0-norm (i.e., $\|\mathbf{z}\|_0 = \sum_i |z_i|^0$) to achieve the sparsity by solving the corresponding relaxed problems, where z_i is the i-th element of vector $\mathbf{z}$. The low-rankness of a matrix is an extension of the vector sparsity concept, which is defined as the number of non-zero singular values of the matrix. A widely adopted approach is to relax the rank function as the nuclear norm, which is the sum of the singular values of the matrix (i.e., the ℓ_1-norm of the singular vector).

Unfortunately, many existing works solve the recovery problem by using the relaxed ℓ_1-norm, which frequently performs less satisfactory in many real-world applications. Although much existing work has theoretically shown that the solution of the relaxed ℓ_1-norm minimization is the same as the initial ℓ_0-norm minimization under certain incoherence conditions [14], [15], the ℓ_1-minimization may be sub-optimal in practice as the incoherence conditions are often too strong to be satisfied in many real-world applications. To address this issue, researchers have proposed several non-convex surrogate functions of the ℓ_0-norm, such as ℓ_p-norm ($0 < p < 1$) [16], Geman [17], Laplace [18], LOG norm [19], Logarithm [20], and ETP [21], to bridge the gap between the ℓ_0-norm and the ℓ_1-norm. These non-convex surrogate functions have achieved better performance than

- *Xiaoqin Zhang, Jingjing Zheng, and Guiying Tang are with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, Zhejiang 325035, China. E-mail: {zhangxiaoqinnan, jjzheng233}@gmail.com, guiyingtang9503@163.com.*
- *Di Wang is with the Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China. E-mail: wang.di@xjtu.edu.cn.*
- *Zhengyuan Zhou is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. E-mail: zyzhou@stanford.edu.*
- *Zhouchen Lin is with the Key Laboratory of Machine Perception (MOE), School of Intelligence Science and Technology, Peking University, Beijing 100871, China. E-mail: zlin@pku.edu.cn.*

Manuscript received 19 November 2021; revised 4 September 2022; accepted 4 October 2022. Date of publication 11 October 2022; date of current version 3 April 2023.
This work was supported in part by the National Natural Science Foundation of China under Grants 61922064, U2033210, 62101387, and 62276004 and in part by the major key project of PCL under Grant PCL2021A12.
(Corresponding author: Xiaoqin Zhang.)
Recommended for acceptance by M. Salzmann.
This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2022.3213716, provided by the authors.
Digital Object Identifier no. 10.1109/TPAMI.2022.3213716

the convex models based on ℓ_1-minimization in signal recovery [22] and classification tasks [23], [24]. In a similar vein, non-convex surrogate functions have been used to improve performance on low-rank recovery problems [22], [25], [26], [27]. When a non-convex surrogate is applied to sparse representation or low-rank recovery problems, the corresponding optimization algorithms have been thoroughly investigated [25], [26], [27], [28], [29]. For non-convex surrogates which satisfy certain conditions, Lu et al. [27] give a generalized solver with an (asymptotic) linear convergence rate[1]. Sparsity-based approaches have demonstrated their effectiveness in many applications, including image clustering [30], [31], [32], face recognition [33], [34], [35], and tracking [36], [37], [38].

However, seeking the sparsest representation of each sample individually may not be the best criterion because it does not consider the structured relationship hidden in samples. To overcome this drawback, structured sparsity, as an important regularizer, which finds jointly sparse representations, has been proposed and studied in the past decade [8], [39], [40]. The core idea of structured sparsity is the sharing mechanism in which samples in the same class should share some common features, promoting the block clustering of non-zero coefficients of sparse representations so that the intrinsic structured information of samples can be fully explored. At a high level, a typical structured sparsity form is to impose the mixed $\ell_{2,0}$-norm regularization on sparse coefficient matrix Z, i.e., $\|Z\|_{2,0} = \sum_i \|\mathbf{z}_i\|_2^0$, where $\mathbf{z}_i$ is the i-th column vector of matrix Z[2]. It is easy to see that the $\ell_{2,0}$-norm is also an extension of the ℓ_0-norm of a vector. A key building block which is essential for solving the structured sparsity problems can be formulated as:

$$\arg\min_Z \frac{1}{2}\|Y - Z\|_F^2 + \lambda\|Z\|_{2,0}, \tag{1}$$

where $Y, Z \in \mathbb{R}^{m\times n}$. As the $\ell_{2,0}$-norm is discontinuous and non-convex, the above problem is NP-hard. Similar to the way for treating the ℓ_0-norm, $\ell_{2,0}$-norm is usually relaxed to $\ell_{2,1}$-norm, and thus problem (1) becomes

$$\arg\min_Z \frac{1}{2}\|Y - Z\|_F^2 + \lambda\|Z\|_{2,1}, \tag{2}$$

where the $\ell_{2,1}$-norm is defined as $\|Z\|_{2,1} = \sum_{i=1}^n \|\mathbf{z}_i\|_2$, i.e., the ℓ_1-norm of the vector with elements being the ℓ_2-norm of the columns of Z. The closed-form solution of the relaxed convex problem (2) is studied in [42], [43]. Unfortunately, the problem (2) suffers from a sub-optimality phenomenon similar to and sometimes even worse than the big gap between the ℓ_0-norm and the ℓ_1-norm for vector sparsity problems. The non-convex surrogate $\ell_{2,p}$-norm, i.e., $\|Z\|_{2,p} = (\sum_i \|\mathbf{z}^i\|_2^p)^{\frac{1}{p}}$ $(0 < p < 1)$ [41], was proposed to provide an improved structured sparsity surrogate. Subsequently, Wang et al. [24] gave a solver for the surrogate $\ell_{2,p}$-norm. But the general solver of structured sparsity for other non-convex surrogate functions is not well studied, and the performance of different surrogate functions remains unclear.

To sum up, there are two major challenges of sparse recovery: (1) the first is how to select the surrogate functions for the $\ell_{2,0}$-norm based structured sparse objectives. Therefore, in order to investigate the performance of various surrogate functions in structured sparse, a unified algorithmic framework with a guarantee of convergence for a wide range of surrogate functions is required. (2) Another issue is the bottleneck of convergence speed in generalized solvers, which affects not only structured sparsity but also vector sparsity and low-rankness. To deal with large-scale data, an efficient general solver for sparse recovery with non-convex surrogates is necessary.

The aim of this work is to address the above issues, and perform both accurate and efficient sparse recovery that can significantly enhance the performance of subsequent applications. Our contributions are four-fold as follows:

- First, we present a general optimization framework (Algorithm 1) that leverages structured sparsity to achieve superior recovery results. A key feature of our framework is that we move significantly beyond the convex and analytically computable $\ell_{2,1}$-norm surrogate objective and allow for a wide range of (possibly non-convex) surrogate functions that serve as better proxies of the computationally infeasible $\ell_{2,1}$-norm based objectives.
- Second, we develop a novel iterative scheme (Algorithm 2), with the global optimal solution being the fixed point of the underlying designed equation[3]. We then establish that the designed fixed point iteration has a global contractive property and converges at an (asymptotic) super-linear rate to a globally optimal solution, despite the fact that the underlying optimization problem is non-convex[4]. The theoretical analysis of the proposed iterative scheme is given in Sections 3.2 and 3.3. Simulation results in Section 5.1 further verify the fast empirical convergence rate. The proposed iterative scheme helps to give a high-efficiency generalized solver for sparse recovery with non-convex surrogates. We apply the iterative scheme to the proposed optimization framework to guarantee the accuracy and efficiency of the framework.
- Third, to fully illustrate the wide applicability of structured sparsity and demonstrate the generality of our framework, in Section 4, we present three concrete problems (i.e., outlier pursuit, supervised feature selection, and structured dictionary learning) in which structured sparsity can be used to formulate meaningful objectives and the corresponding relaxed

1. To improve the efficacy of the generalized solver, we present a novel generalized solver with an (asymptotic) superlinear convergence rate in this paper. Both theoretically and experimentally, we show that the proposed solver's convergence rate is much faster than Lu's work [27].

2. In some works [24], [41], $\ell_{2,0}$-norm is defined on the rows of matrix Z as $\|Z\|_{2,0} = \sum_i \|\mathbf{z}^i\|_2^0$, where $\mathbf{z}^i$ is the i-th row vector of the matrix Z. This does not affect the main results of this paper because the two definitions of structured sparsity are the same with the transpose operator.

3. A basic idea of a fixed-point algorithm is to build an iteration function $\mathcal{J}(x)$ such that the sequence obtained by $x_{n+1} = \mathcal{J}(x_n)$, $n = 0, 1, 2, \ldots$ converges to an optimal solution of the optimization problem.

4. If the objective function of an optimization problem is non-convex, or the feasible set of the optimization problem is non-convex, the optimization problem is a non-convex optimization problem.

formulations using (possibly non-convex) surrogates can be solved by the proposed general optimization framework (Algorithm 1).

- Fourth, to demonstrate the efficacy of our framework, we conduct extensive experiments using both synthetic and real-world data. In particular, on the real-world data front, we evaluate our proposed framework with three applications, including outlier pursuit, supervised feature selection, and structured dictionary learning. The empirical performance of the recovered solutions demonstrates that ℓ_1-norm based relaxation usually provides worse recovery quality compared to the non-convex surrogates. A second insight is that the best performing non-convex surrogate(s) for different applications are frequently different. Nonetheless, we do not need to design a specific algorithm for each such surrogate function from scratch because our framework provides a unified approach from a computational standpoint. We consider this computation/optimization framework that can simultaneously handle a wide set of surrogate functions (which have different strengths in different applications). This is the key contribution of our work, both from a theoretical and an applied standpoint.

The rest of this paper is structured as follows. Section 2 is the problem formulation. Some notations and the main work of this paper are introduced. The proof related to the convergence analysis of the proposed algorithms is given in Section 3. In Section 4, three concrete applications of the proposed algorithms are discussed in detail. Section 5 presents numerical experiments conducted on synthetic and real data. We conclude this work in Section 6.

TABLE 1
A Summary of the Definitions and Symbols Used in This Paper

Notations	Meanings
$\mathbf{0}$	Null vector.
x_i	The i-th element of vector $\mathbf{x}$.
$\|\mathbf{x}\|_0$	ℓ_0 norm of vector $\mathbf{x}$, i.e., number of nonzero entries of $\mathbf{x}$.
$\|\mathbf{x}\|_1$	ℓ_1 norm of vector $\mathbf{x}$, , i.e., $\|\mathbf{x}\|_1 = \sum_{i=1}^m \|x_i\|$, where x_i is the i-th element of $\mathbf{x}$.
$\|\mathbf{x}\|_p$	ℓ_p norm of vector $\mathbf{x}$, , i.e., $\|\mathbf{x}\|_p = (\sum_{i=1}^m \|x_i\|^p)^{\frac{1}{p}}$.
$\langle \mathbf{x}, \mathbf{y} \rangle$	The inner product of $\mathbf{x}$ and $\mathbf{y}$, , i.e., $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^m x_i y_i$.
x_{ij}	Each element in X is represented as x_{ij}.
$\mathbf{x}_i$ and $\mathbf{x}^i$	Each column in X is represented as $\mathbf{x}_i$, and each row in X is represented as $\mathbf{x}^i$.
$\|X\|_0$	ℓ_0-norm of X, i.e., number of nonzero entries of X.
$\|X\|_F$	Frobenius norm of X, i.e., $\|X\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n x_{ij}^2}$.
$\|X\|_1$	ℓ_1-norm of X, i.e., $\|X\|_1 = \sum_{i=1}^m \sum_{j=1}^n \|x_{ij}\|$.
$\|X\|_p$	ℓ_p-norm of X, , i.e., $\|X\|_p = (\sum_{i=1}^m \sum_{j=1}^n \|x_{ij}\|^p)^{\frac{1}{p}}$.
$\|X\|_{2,0}$	$\ell_{2,0}$-norm of X, , i.e., number of nonzero column vectors of X.
$\|X\|_{2,1}$	$\ell_{2,1}$-norm of X, i.e., $\|X\|_{2,1} = \sum_{j=1}^n \|\mathbf{x}_j\|_2$.
$\|X\|_{2,p}$	$\ell_{2,p}$-norm of X, i.e., $\|X\|_{2,p} = (\sum_{j=1}^n \|\mathbf{x}_j\|^p)^{\frac{1}{p}}$.
$\sigma(X)$	The vector composed of singular values of matrix X.
$\sigma_i(X)$	The i-th singular value of matrix X.
$\mathrm{rank}(X)$	Rank function of matrix X, i.e., number of nonzero entries of singular values.
$\|X\|_*$	Nuclear norm of X, i.e., $\sum_{i=1}^{\mathrm{rank}(X)} \sigma_i(X)$.
$\langle X, Y \rangle$	The inner product of X and Y, , i.e., $\langle X, Y \rangle = \sum_{i=1}^m \sum_{j=1}^n x_{ij} y_{ij}$.
$\mathrm{Diag}(\mathbf{x})$	A diagonal matrix with the i-th diagonal element being x_i.

2 Problem Formulation

For clarity, we proceed with the following notational conventions: 1) Lowercase letters for scalars ($x, y, z \cdots$); 2) Bold lowercase letters for vectors ($\mathbf{x}, \mathbf{y}, \mathbf{z} \cdots$); 3) Capital letters for matrices ($X, Y, Z \cdots$); 4) Calligraphic letters for functions ($\mathcal{X}, \mathcal{Y}, \mathcal{Z} \cdots$); 5) Swash letters for sets ($\mathscr{X}, \mathscr{Y}, \mathscr{Z} \cdots$). Specifically, $\mathbb{R}$ stands for fields of real numbers, and $\emptyset$ stands for empty set. More definitions and symbols related to vectors and matrices are given in Table 1.

In this paper, we study a family of non-convex surrogate functions of $\ell_{2,0}$-norm and fill in the gap between the $\ell_{2,0}$-norm and the $\ell_{2,1}$-norm. A general non-convex approximation of problem (1) can be formulated as:

$$\mathcal{S}(Y, \lambda) = \arg\min_Z \frac{1}{2}\|Y - Z\|_F^2 + \lambda \sum_{i=1}^n \mathcal{G}(\|\mathbf{z}_i\|_2), \tag{3}$$

where the surrogate function $\mathcal{G}(\cdot) : [0, +\infty) \to [0, +\infty)$ satisfies:

- **A1** $\mathcal{G}(x)$ is strictly concave and increasing, and $\mathcal{G}(0) = 0$;
- **A2** $\mathcal{G}'(x)$ is strictly convex;
- **A3** $\mathcal{G}''(x)$ is continuous on $(0, +\infty)$.

Our assumptions about the surrogate function $\mathcal{G}(\cdot)$ are easy to satisfy and include many non-convex surrogate functions in the literature. For a sample list, see Table 2. Hence, $\sum_{i=1}^n \mathcal{G}(\|\mathbf{z}_i\|_2)$ can be considered as a general non-convex surrogate of the $\ell_{2,0}$-norm, and it is expected to have a better approximation of the $\ell_{2,0}$-norm than the convex relaxation $\ell_{2,1}$-norm. Note that the structured sparsity optimization problem (3), of course, includes vector sparsity as a simple special case. But even in this important special case, only individual non-convex surrogates have been studied on a case-by-case basis, and there has not been a framework that deals with a wide class of (non-convex) surrogate functions simultaneously with global provable convergence guarantees.

Algorithm 1. Structured Sparsity Optimization Framework for Solving (3)

Input: $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \in \mathbb{R}^{m \times n}$, a threshold $\lambda > 0$
Output: $\mathcal{S}(Y, \lambda)$ as defined in (3).
for $i = 1, 2, \ldots n$ **do**
  if $\mathbf{y}_i = \mathbf{0}$ **then**
    $\mathscr{A}_i = \{\mathbf{0}\}$.
  else
    $\mathscr{A}_i = \{x^* \frac{\mathbf{y}_i}{\|\mathbf{y}_i\|} | x^* \in \mathscr{B}_i\}$, where $\mathscr{B}_i = \mathrm{Solve}(\|\mathbf{y}_i\|_2, \lambda)$.
  end
end
Let $\mathcal{S}(Y, \lambda) = [\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*]$, where $\mathbf{x}_i^* \in \mathscr{A}_i$ for $i = 1, 2, \ldots, n$.

In the rest of this paper, to solve (3) more accurately and efficiently, we study the key optimization problem (3) in

TABLE 2
Examples of Surrogate Functions of ℓ_0, Where $\gamma > 0$, and a_0 is Defined by Eq. (18)

Name	$\mathcal{G}(x)$	$\mathcal{G}'(x)$	$\mathcal{G}''(x)$	a_0
ℓ_p [16]	$x^p, 0 < p < 1$	px^{p-1}	$p(p-1)x^{(p-2)}$	$\max((\lambda p(1-p))^{\frac{1}{2-p}}, 0)$
Geman [17]	$\frac{x}{x+\gamma}$	$\frac{\gamma}{(\gamma+x)^2}$	$\frac{-2\gamma}{(\gamma+x)^3}$	$\max((2\lambda\gamma)^{\frac{1}{3}} - \gamma, 0)$
Laplace [18]	$(1-\exp(-\frac{x}{\gamma}))$	$\frac{1}{\gamma}\exp(\frac{-x}{\gamma})$	$-\frac{1}{\gamma^2}\exp(\frac{-x}{\gamma})$	$\max(-\gamma\log(\frac{\gamma^2}{\lambda}), 0)$
LOG [19]	$\log(\gamma + x)$	$\frac{1}{\gamma+x}$	$-\frac{1}{(\gamma+x)^2}$	$\max(\sqrt{\lambda} - \gamma, 0)$
Logarithm [20]	$\frac{1}{\log(\gamma+1)}\log(\gamma x + 1)$	$\frac{\gamma}{(\gamma x+1)\log(\gamma+1)}$	$-\frac{\gamma^2}{(\gamma x+1)^2\log(\gamma+1)}$	$\max\left(\frac{\sqrt{\frac{\gamma^2\lambda}{\log(\gamma+1)}}-1}{\gamma}, 0\right)$
ETP [21]	$\frac{1-\exp(-\gamma x)}{1-\exp(-\gamma)}$	$\frac{\gamma\exp(-\gamma x)}{1-\exp(-\gamma)}$	$\frac{-\gamma^2\exp(-\gamma x)}{1-\exp(-\gamma)}$	$\max\left(\frac{\log(\frac{1-\exp(-\gamma)}{\lambda\gamma^2})}{-\gamma}, 0\right)$

depth and design a novel optimization algorithm (Algorithm 1) that solves this problem to global optimality efficiently. More specifically, our algorithm hinges on developing a novel iterative scheme to (4) (Algorithm 2), with the global optimal solution being the fixed point of the underlying designed equation (5).

$$\arg\min_{x\geq 0} \mathcal{F}_y(x) = \frac{1}{2}(y-x)^2 + \lambda\mathcal{G}(x). \tag{4}$$

$$\begin{cases} \mathcal{J}_1(x) = y - \lambda\mathcal{G}'(x), \\ \mathcal{J}_2(x) = \mathcal{J}_1(x) - \frac{(\mathcal{J}_1(\mathcal{J}_1(x))-\mathcal{J}_1(x))(\mathcal{J}_1(x)-x)}{\mathcal{J}_1(\mathcal{J}_1(x))-2\mathcal{J}_1(x)+x}. \end{cases} \tag{5}$$

By [27], for any lower bounded function $\mathcal{G}: [0, +\infty) \to [0, +\infty)$, the optimal solution to

$$\mathcal{T}(Y, \lambda) = \arg\min_X \frac{1}{2}\|Y - X\|_F^2 + \lambda \sum_{i=1}^{\min(m,n)} \mathcal{G}(\sigma_i(X)) \tag{6}$$

can be calculated by $X^* = U\text{Diag}(\boldsymbol{\sigma}^*)V^T$, where U and V are obtained from the SVD of $Y \in \mathbb{R}^{m\times n}$: $Y = U\text{Diag}(\sigma(Y))V^T$, and the i-th element of $\boldsymbol{\sigma}^*$ is

$$\sigma_i^* \in \text{Prox}_{\lambda\mathcal{G}}(\sigma_i(Y)) = \arg\min_{x\geq 0} \frac{1}{2}(\sigma_i(Y) - x)^2 + \lambda\mathcal{G}(x).$$

As a result, we have Algorithm 3 for solving (6) efficiently.

For convenience, we term Algorithms 2 and 3 as Generalized Accelerating Iterative (GAI) and Generalized Singular Value Thresholding by GAI (GSVT-GAI) respectively for the remainder of the paper. We prove the proposed three algorithms have a global contractive property and converge at a linear rate to a globally optimal solution. Furthermore, asymptotically, the proposed algorithms converge even faster than Lu's work, at a super-linear rate.

3 A Novel Optimization Framework for Structured Sparsity

In the following subsections, a novel optimization framework (Algorithm 1) is proposed to solve the non-convex optimization problem (3). We first characterize the solution set of the optimization problem (3) in terms of a simpler one-dimensional non-convex optimization problem (4), and then design a novel fixed point iteration scheme (GAI) to solve (4). Finally, we show that the proposed algorithm converges to a global optimal solution at a linear rate. Furthermore, asymptotically, the algorithm converges to a global optimal solution at a faster, super-linear rate.

Algorithm 2. Generalized Accelerating Iterative Algorithm (GAI) for Solving (4)

Input: A real number $y > 0$, a threshold $\lambda > 0$, and a tolerance $\tau > 0$.
Output: $\text{Solve}(y, \lambda) = x_{\text{G}}^*$.
$a_0 \leftarrow \max\{x | \mathcal{J}_1'(x) = 1 \text{ or } x = 0\}$.
Let

$$\begin{cases} \mathcal{F}_y(x) = \frac{1}{2}(y-x)^2 + \lambda\mathcal{G}(x) \\ \mathcal{J}_1(x) = y - \lambda\mathcal{G}'(x), \\ \mathcal{J}_2(x) = \mathcal{J}_1(x) - \frac{(\mathcal{J}_1(\mathcal{J}_1(x))-\mathcal{J}_1(x))(\mathcal{J}_1(x)-x)}{\mathcal{J}_1(\mathcal{J}_1(x))-2\mathcal{J}_1(x)+x} \end{cases},$$

if $\mathcal{F}_y'(a_0) < 0$ **then**
  // Find $\hat{x}_{\text{G}}$ by fixed point iteration
  Initialize $x_{\text{G}}^{(0)} \leftarrow y$
  $k \leftarrow 0$
  while $|\mathcal{J}_1(\mathcal{J}_1(x_{\text{G}}^{(k)})) - 2\mathcal{J}_1(x_{\text{G}}^{(k)}) + x_{\text{G}}^{(k)}| > \tau$ **do**
    $x_{\text{G}}^{(k+1)} = \mathcal{J}_2(x_{\text{G}}^{(k)})$
    $k \leftarrow k + 1$
  end
  $\hat{x}_{\text{G}} = \mathcal{J}_1(x_{\text{G}}^{(k)})$
else
  return $\hat{x}_{\text{G}} = a_0$
end
If $\mathcal{F}_y(0) > \mathcal{F}_y(\hat{x}_{\text{G}})$, return $x_{\text{G}}^* = \hat{x}_{\text{G}}$; otherwise return $x_{\text{G}}^* = 0$.

3.1 Solution Set Characterization

We start by characterizing the solution set of the non-convex optimization problem (3). Since the original objective function of problem (3) can be easily decomposed as

$$\arg\min_Z \sum_{i=1}^{n}\left\{\frac{1}{2}\|\mathbf{y}_i - \mathbf{z}_i\|_2^2 + \lambda\mathcal{G}(\|\mathbf{z}_i\|_2)\right\},$$

it suffices to consider the following subproblem

$$\arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{y} - \mathbf{z}\|_2^2 + \lambda\mathcal{G}(\|\mathbf{z}\|_2). \tag{7}$$

To this end, we consider the solution set of a more general formulation as

$$\arg\min_{\mathbf{z}} \mathcal{H}(\mathbf{z}) = \frac{1}{p}\|\mathbf{y}-\mathbf{z}\|_p^p + \lambda\mathcal{G}(\|\mathbf{z}\|_p), \tag{8}$$

where $p \geq 1$, and $\|\mathbf{z}\|_p = (\sum_{i=1} |z_i|^p)^{\frac{1}{p}}$.

Note that when $\mathbf{y} = \mathbf{0}$, the unique optimal solution to problem (8) is $\mathbf{0}$. Hence, we only need to consider the solution set of problem (8) when $\mathbf{y} \neq \mathbf{0}$. To facilitate the exposition of the solution set characterization, we define two functions for a given vector $\mathbf{y}$,

$$\mathcal{H}(\mathbf{z}) = \frac{1}{p}\|\mathbf{y}-\mathbf{z}\|_p^p + \lambda\mathcal{G}(\|\mathbf{z}\|_p),$$

$$\mathcal{F}(x) = \frac{1}{p}\Big|\|\mathbf{y}\|_p - x\Big|^p + \lambda\mathcal{G}(|x|),$$

as well as two corresponding sets

$$\mathscr{A} = \arg\min\nolimits_{\mathbf{z}} \mathcal{H}(\mathbf{z}), \quad \mathscr{B} = \left\{ x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p} \middle| x^* \in \arg\min\nolimits_x \mathcal{F}(x) \right\}.$$

Note that $\mathscr{A}$ is the solution set that we wish to obtain. The next lemma gives a full characterization of the relationship between the sets $\mathscr{A}$ and $\mathscr{B}$.

Algorithm 3. Generalized Singular Value Thresholding by GAI (GSVT-GAI) for solving (6)

Input: Y, λ and a tolerance $\tau > 0$.
Output: $\mathcal{T}(Y,\lambda)$.
$a_0 \leftarrow \max\{x | \lambda\mathcal{G}'(x) = -1 \text{ or } x = 0\}$;
$[U, \sigma(Y), V] = \text{svd}(Y)$; $i = 0$;
while $i \leq \text{length}(\sigma(Y))$ **do**
  $i = i + 1$;
  if $\sigma_i(Y) = 0$ **then**
    $\hat{\sigma}_i = 0$; Break;
  end
  if $\sigma_i(Y) > \lambda\mathcal{G}'(a_0) + a_0$ **then**
    // Find $\hat{x}_{\text{G}}$ by fixed point iteration
    Initialize $x_{\text{G}}^{(0)} \leftarrow \sigma_i(Y)$; $k \leftarrow 0$
    Let $\begin{cases} \mathcal{J}_1(x) = \sigma_i(Y) - \lambda\mathcal{G}'(x), \\ \mathcal{J}_2(x) = \mathcal{J}_1(x) - \frac{(\mathcal{J}_1(\mathcal{J}_1(x)) - \mathcal{J}_1(x))(\mathcal{J}_1(x) - x)}{\mathcal{J}_1(\mathcal{J}_1(x)) - 2\mathcal{J}_1(x) + x} \end{cases}$.
    while $|\mathcal{J}_1(\mathcal{J}_1(x_{\text{G}}^{(k)})) - 2\mathcal{J}_1(x_{\text{G}}^{(k)}) + x_{\text{G}}^{(k)}| > \tau$ **do**
      $x_{\text{G}}^{(k+1)} = \mathcal{J}_2(x_{\text{G}}^{(k)})$; $k \leftarrow k + 1$
    end
    $\hat{x}_{\text{G}} = \mathcal{J}_1(x_{\text{G}}^{(k)})$
  else
    $\hat{x}_{\text{G}} = a_0$
  end
  If $0 > \frac{1}{2}\hat{x}_{\text{G}}^2 - \sigma_i(Y)\hat{x}_{\text{G}} + \lambda\mathcal{G}(\hat{x}_{\text{G}})$, $\hat{\sigma}_i = \hat{x}_{\text{G}}$; otherwise $\hat{\sigma}_i = 0$;
  if $\hat{\sigma}_i = 0$ **then**
    Break;
  end
end
Compute $\mathcal{T}(Y,\lambda) = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_i]\text{Diag}(\hat{\sigma})[\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_i]^T$.

Lemma 1. *If $p \geq 1$, $\mathcal{G}(\cdot)$ is continuous on $[0,+\infty)$, $\mathcal{G}(|x|) \geq 0$ for $x \in \mathbb{R}$, and $\mathcal{G}(x) = 0$ if and only if $x = 0$, then*

$$\mathscr{A} = \begin{cases} \{\mathbf{0}\}, & \text{if } \mathbf{y} = \mathbf{0}, \\ \mathscr{B}, & \text{if } \mathbf{y} \neq \mathbf{0}. \end{cases}$$

Proof. The proof is broken into three steps.

Step 1: The existence of optimal solution to $\arg\min_x \mathcal{F}(x)$. Note that for $\forall\, x < 0$, we have

$$\mathcal{F}(x) = \frac{1}{p}\Big|\|\mathbf{y}\|_p - x\Big|^p + \lambda\mathcal{G}(|x|) > \frac{1}{p}\|\mathbf{y}\|_p^p + \lambda\mathcal{G}(0) = \mathcal{F}(0),$$

which is followed from the fact that $\frac{1}{p}|\|\mathbf{y}\|_p - x|^p > \frac{1}{p}\|\mathbf{y}\|_p^p$ for $\forall\, x < 0$ and $\lambda\mathcal{G}(|x|) \geq 0 = \lambda\mathcal{G}(0)$. For $\forall\, x > 2\|\mathbf{y}\|_p$, because $\frac{1}{p}|\|\mathbf{y}\|_p - x|^p > \frac{1}{p}\|\mathbf{y}\|_p^p$, we have

$$\mathcal{F}(x) > \frac{1}{p}\|\mathbf{y}\|_p^p + \lambda\mathcal{G}(|x|) \geq \frac{1}{p}\|\mathbf{y}\|_p^p + \lambda\mathcal{G}(0) = \mathcal{F}(0).$$

Since $\mathcal{F}(x)$ is lower bounded from $\mathcal{F}(x) \geq 0$, the infimum of $\mathcal{F}(x)$ (denoted by $\inf_x \mathcal{F}(x)$) exists. Thus, we obtain

$$\inf_x \mathcal{F}(x) = \inf_{x \in [0, 2\|\mathbf{y}\|_p]} \mathcal{F}(x).$$

As $\mathcal{G}(\cdot)$ is continuous on $[0,+\infty)$, resulting $\mathcal{F}(x)$ is continuous on $[0, 2\|\mathbf{y}\|_p]$. Thus $\mathcal{F}(x)$ has a minimum on $[0, 2\|\mathbf{y}\|_p]$. Mark the minimum of $\mathcal{F}(x)$ for $x \in [0, 2\|\mathbf{y}\|_p]$ as $v_{\mathcal{F}}$. By intermediate value theorem, there exists $x^* \in [0, 2\|\mathbf{y}\|_p]$ such that $\mathcal{F}(x^*) = v_{\mathcal{F}}$, i.e.,

$$\inf_x \mathcal{F}(x) = \mathcal{F}(x^*).$$

Therefore, x^* is a minimizer of $\mathcal{F}(x)$, and $\mathscr{B} \neq \emptyset$.

Step 2: Identification of infima. Since $\mathcal{H}(\mathbf{z}) \geq 0$ for any $\mathbf{z} \in \mathbb{R}^m$, $\inf_{\mathbf{z}} \mathcal{H}(\mathbf{z})$ exists. The infimum of $\mathcal{H}(\mathbf{z})$ for $\mathbf{z} \in \mathbb{R}^m$ is denoted by $v_{\mathcal{H}}$, i.e., $v_{\mathcal{H}} = \inf_{\mathbf{z}} \mathcal{H}(\mathbf{z})$. We prove $v_{\mathcal{H}} = v_{\mathcal{F}}$ for $\forall\, \mathbf{y} \neq \mathbf{0}$ in this step. Based on the results of Step 1, for $\forall\, \mathbf{z} \in \mathbb{R}^m$, we have

$$\begin{aligned} v_{\mathcal{F}} = \mathcal{F}(x^*) \leq \mathcal{F}(\|\mathbf{z}\|_p) &= \frac{1}{p}|\|\mathbf{y}\|_p - \|\mathbf{z}\|_p|^p + \lambda\mathcal{G}(\|\mathbf{z}\|_p) \\ &\leq \frac{1}{p}\|\mathbf{y}-\mathbf{z}\|_p^p + \lambda\mathcal{G}(\|\mathbf{z}\|_p) = \mathcal{H}(\mathbf{z}), \end{aligned} \tag{9}$$

where the forth inequality is derived from the Minkowski inequality.[5] Taking the infimum over all $\mathbf{z}$ for the function $\mathcal{H}(\mathbf{z})$, we conclude from (9) that $v_{\mathcal{F}} \leq v_{\mathcal{H}}$. On the other hand,

$$\begin{aligned} v_{\mathcal{H}} &\leq \mathcal{H}\left(x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p}\right) \\ &= \frac{1}{p}\left\|\mathbf{y} - x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p}\right\|_p^p + \lambda\mathcal{G}\left(\left\|x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p}\right\|_p\right) \\ &= \frac{1}{p}\left\|(\|\mathbf{y}\|_p - x^*)\frac{\mathbf{y}}{\|\mathbf{y}\|_p}\right\|_p^p + \lambda\mathcal{G}(|x^*|) \\ &= \frac{1}{p}|\|\mathbf{y}\|_p - x^*|^p + \lambda\mathcal{G}(|x^*|) = v_{\mathcal{F}}. \end{aligned}$$

5. By Minkowski inequality, $|\|\mathbf{y}\|_p - \|\mathbf{z}^*\|_p| = \|\mathbf{y} - \mathbf{z}^*\|_p$ holds if and only if there exists $a \geq 0$ such that $\mathbf{y} = a\mathbf{z}^*$ or $\mathbf{z}^* = a\mathbf{y}$. If $\mathbf{y} = a\mathbf{z}^*$, since $\mathbf{y} \neq \mathbf{0}$ and $a \neq 0$, we set c_0 as $\frac{\|\mathbf{y}\|_p}{a}$. If $\mathbf{z}^* = a\mathbf{y}$, we set c_0 as $a\|\mathbf{y}\|_p$.

Thus, there holds

$$v_{\mathcal{F}} = v_{\mathcal{H}} = \mathcal{H}\left(x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p}\right), \tag{10}$$

from which we know that $x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p}$ is a minimizer of $\mathcal{H}(\mathbf{z})$ and $\mathscr{A} \neq \emptyset$.

Step 3: Identification of the solution set. It is clear that $\mathscr{A} = \mathscr{B} = \{\mathbf{0}\}$ when $\mathbf{y} = \mathbf{0}$. Therefore, we only need to prove that $\mathscr{A} = \mathscr{B}$ when $\mathbf{y} \neq \mathbf{0}$ in the following. On one hand, for $\forall\ x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p} \in \mathscr{B}$, it can be easily concluded from (10) that $x^* \frac{\mathbf{y}}{\|\mathbf{y}\|_p} \in \mathscr{A}$. Hence, $\mathscr{B} \subseteq \mathscr{A}$.

On the other hand, let $\mathbf{z}^* \in \mathscr{A}$. The following must be observed

$$\begin{aligned} v_{\mathcal{H}} = v_{\mathcal{F}} &\leq \frac{1}{p}|\|\mathbf{y}\|_p - \|\mathbf{z}^*\|_p|^p + \lambda\mathcal{G}(\|\mathbf{z}^*\|_p) \\ &\leq \frac{1}{p}\|\mathbf{y} - \mathbf{z}^*\|_p^p + \lambda\mathcal{G}(\|\mathbf{z}^*\|_p) = v_{\mathcal{H}}, \end{aligned}$$

which indicates that

$$\left|\|\mathbf{y}\|_p - \|\mathbf{z}^*\|_p\right| = \|\mathbf{y} - \mathbf{z}^*\|_p, \tag{11}$$

$$v_{\mathcal{F}} = \frac{1}{p}|\|\mathbf{y}\|_p - \|\mathbf{z}^*\|_p|^p + \lambda\mathcal{G}(\|\mathbf{z}^*\|_p)\ . \tag{12}$$

According to (11), it can be concluded that $\mathbf{z}^* = c_0 \frac{\mathbf{y}}{\|\mathbf{y}\|_p}$ for some $c_0 \geq 0$ by Minkowski inequality. Combining this with (12), we know that c_0 is a minimizer of $\mathcal{F}(x)$. Thus, $\mathbf{z}^* = c_0 \frac{\mathbf{y}}{\|\mathbf{y}\|_p} \in \mathscr{B}$. The proof for the identification of the solution set is complete. □

Algorithm 4. Lu's Work [27]

Input: A real number $y > 0$, a number of iterations $\kappa > 0$ and a tolerance $\tau > 0$.
Output: x_{L}^*.
// Find $\hat{x}_{\mathrm{L}}$ by fixed point iteration
Initialize $x_{\mathrm{L}}^{(0)} = y$, $x_{\mathrm{L}}^{(1)} = \mathcal{J}_1(x_{\mathrm{L}}^{(0)})$ and $k = 1$.
while $|x_{\mathrm{L}}^{(k)} - x_{\mathrm{L}}^{(k-1)}| > \tau$ & $k < \kappa$ **do**
  $x_{\mathrm{L}}^{(k+1)} = \mathcal{J}_1(x_{\mathrm{L}}^{(k)})$.
  if $x_{\mathrm{L}}^{(k+1)} < 0$ **then**
    return $\hat{x}_{\mathrm{L}} = 0$.
  else
    $\hat{x}_{\mathrm{L}} = x_{\mathrm{L}}^{(k+1)}$.
  end
  Let $k = k + 1$.
end
Compare $\mathcal{F}_y(0)$ and $\mathcal{F}_y(\hat{x}_{\mathrm{L}})$ to identify the optimal solution x_{L}^*.

The above lemma then immediately leads to the following conclusion:

Corollary 1. *Suppose* $\mathbf{y} \neq \mathbf{0}$. *Let* $p = 2$, $y = \|\mathbf{y}\|_2$, *and* $\mathscr{C} = \arg\min_{x \geq 0} \frac{1}{2}(y - x)^2 + \lambda\mathcal{G}(x)$. *Then* $\mathscr{A} = \{c\frac{\mathbf{y}}{\|\mathbf{y}\|_2} \,|\, c \in \mathscr{C}\}$.

Proof. It suffices to note that if x^* is a solution of

$$\min_x \frac{1}{2}(y - x)^2 + \lambda\mathcal{G}(|x|) \tag{13}$$

then $x^* \geq 0$. To see this, assume $x^* \in \arg\min_x \frac{1}{2}(y - x)^2 + \lambda\mathcal{G}(|x|)$, then

$$\frac{1}{2}(y - x^*)^2 + \lambda\mathcal{G}(|x^*|) \leq \frac{1}{2}(y - (-x^*))^2 + \lambda\mathcal{G}(|-x^*|)\ ,$$

which implies $x^* y \geq 0$. Therefore, we have $x^* \geq 0$ due to the fact that $y > 0$. □

3.2 A Fixed Point Algorithm

Corollary 1 gives us a convenient solution set characterization for solving the original optimization problem (3), provided we can solve (4). Note that $\mathcal{G}(x)$ is non-convex, and hence the problem (4) is non-convex. In this section, a novel fixed point iterative scheme using the underlying designed equation (5) is developed to achieve this goal.

As follows, we first define some notations for the remainder of this paper.

$$\mathcal{J}_1(x) = y - \lambda\mathcal{G}'(x), \tag{14}$$

$$\mathscr{S} = \{x \,|\, \mathcal{F}_y'(x) = 0, 0 \leq x \leq y\}, \tag{15}$$

$$\bar{x}_y = \max\{x \,|\, x \in \mathscr{S}\}, \tag{16}$$

$$\mathcal{G}''(0) = \lim_{x \to 0+} \mathcal{G}''(x)\ ,\ \mathcal{J}_2(\bar{x}_y) = \lim_{x \to \bar{x}_y^+} \mathcal{J}_2(x), \tag{17}$$

$$a_0 = \max\{x | \mathcal{J}_1'(x) = 1 \text{ or } x = 0\}. \tag{18}$$

The following simple properties are immediate, and the corresponding proofs are given in Appendix, available online.

Property 1. *Given* $\mathcal{G}$ *satisfies Assumptions* ***A1-A3****, we have:*

i) $\mathcal{G}'(x)$ *is strictly decreasing, and* $\mathcal{G}'(x) > 0$ *in* $(0, +\infty)$;
ii) $\mathcal{G}''(x)$ *is strictly increasing, and* $\mathcal{G}''(x) < 0$ *in* $(0, +\infty)$;
iii) $\mathcal{J}_1(x)$ *is strictly increasing, and* $\mathcal{J}_1'(x)$ *is strictly decreasing on* $(0, +\infty)$;
iv) $\mathcal{F}_y(x)$ *is strictly increasing in* $(0, +\infty)$ *if* $\mathscr{S} = \emptyset$.

From Property 1 (iv), we know that the optimal solution to $\arg\min_{x \geq 0} \mathcal{F}_y(x)$ is 0 when $\mathscr{S} = \emptyset$. Thus, the non-trivial case $\mathscr{S} \neq \emptyset$ is considered in the following. We start by providing some intuition. First, a minimizer x^* should either be 0 or satisfy the first-order optimality condition: $\mathcal{F}_y'(x^*) = 0$, i.e., $\mathcal{J}_1(x^*) = x^*$, which implies that a non-zero x^* is a fixed point of $\mathcal{J}_1(x)$. At this point, a natural (and crude) algorithm already suggests itself: start at $x_{\mathrm{L}}^{(0)} = y$, and follow a fixed-point update rule $x_{\mathrm{L}}^{(k+1)} = \mathcal{J}_1(x_{\mathrm{L}}^{(k)})$. However, the above iterate has at most a linear convergence rate, as shown in Theorem 1 (i). To speed up the convergence rate for problem (4), we create a novel iterative function $\mathcal{J}_2(x)$ and propose the iterate $x_{\mathrm{G}}^{(k+1)} = \mathcal{J}_2(x_{\mathrm{G}}^{(k)})$ with a super-linear convergence rate, as shown in Theorem 1 (ii).

To be a complete algorithm, we still need to specify a stopping criterion. The one that we will use here is to check the condition $|\mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)})) - 2\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) + x_{\mathrm{G}}^{(k)}| > \tau$, where $\tau > 0$ is a pre-specified tolerance threshold. To explain the intuition behind this specific choice of the stopping criterion, we next characterize some simple properties of the underlying mathematical objects.

Proposition 1. *[27] Given* $\mathcal{G}$ *satisfies Assumptions* ***A1-A3****. If* $\mathscr{S} \neq \emptyset$, *then:*

i) *The sequence* $\{x_{\mathrm{L}}^{(k)}\}$ *generated by* $x_{\mathrm{L}}^{(k+1)} = \mathcal{J}_1(x_{\mathrm{L}}^{(k)})$ *with the initialization* $x_{\mathrm{L}}^{(0)} = y$, *converges to* $\bar{x}_y$;
ii) $\bar{x}_y < \mathcal{J}_1(x) < x$ *for* $\forall\, x \in (\bar{x}_y, y]$;
iii) $\bar{x}_y$ *is a fixed point of* $\mathcal{J}_1(x)$, *i.e.,* $\mathcal{J}_1(\bar{x}_y) = \bar{x}_y$;
iv) *The global optimal solution to the problem (4) is in the set* $\{0, \bar{x}_y\}$.

Property 2. *Given* $\mathcal{G}$ *satisfies Assumptions* ***A1-A3****. If* $\mathscr{S} \neq \emptyset$, *then:*

i) $0 \leq a_0 < y$;
ii) $\mathcal{J}_1'(a_0) \leq 1$, $\mathcal{J}_1'(\bar{x}_y) \leq 1$, *and* $0 < \mathcal{J}_1'(x) < 1$ *for* $\forall\, x \in (a_0, +\infty)$;
iii) *The equation* $\mathcal{F}_y'(x) = 0$ *has a unique solution in* (a_0, y) *when* $\mathcal{F}_y'(a_0) < 0$;
iv) $\bar{x}_y \notin (0, a_0) \cup (a_0, y]$ *when* $\mathcal{F}_y'(a_0) \geq 0$;
v) *There exists at most one non-zero local minimum to* $\mathcal{F}_y(x)$ *in* $(0, +\infty)$.

From Property 2 (iv) and (v), we know that the optimal solution to the problem (4) should either be 0 or a_0 when $\mathcal{F}_y'(a_0) \geq 0$. Thus, we only consider the case for $\mathcal{F}_y'(a_0) < 0$ (the iterations in Algorithm 2). From Property 2 (iii) and the definition (16) of $\bar{x}_y$, we have $\bar{x}_y > a_0$ when $\mathcal{F}_y'(a_0) < 0$. Now, to see the intuition behind the stopping criterion, we can conclude from the mean value theorem that

$$\begin{aligned}&\left|\mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)})) - 2\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) + x_{\mathrm{G}}^{(k)}\right| \\ &= \left(\frac{1}{\mathcal{J}_1'(\xi)} - 1\right)\left|\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) - \mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)}))\right|,\end{aligned}$$

where $\xi \in (\mathcal{J}_1(x_{\mathrm{G}}^{(k)}), x_{\mathrm{G}}^{(k)})$, and $\frac{1}{\mathcal{J}_1'(\xi)} - 1 > 0$ follows from Proposition 1 (ii) and Property 2 (ii). Since $\bar{x}_y > a_0$, we know from Property 2 (ii) that $0 < \mathcal{J}_1'(\bar{x}_y) < 1$. Let $c_0 = \frac{1}{\frac{1}{\mathcal{J}_1'(\bar{x}_y)} - 1}$, we can obtain from Property 1 (iii) that

$$\begin{aligned}&c_0\left|\mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)})) - 2\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) + x_{\mathrm{G}}^{(k)}\right| \\ &\geq \left|\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) - \mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)}))\right|.\end{aligned}$$

Therefore, there are two merits to set $\left|\mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)})) - 2\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) + x_{\mathrm{G}}^{(k)}\right| < \tau$ as a stopping criterion of Algorithm 2 as follows.

- The value of

$$\left|\mathcal{F}_y'(\hat{x}_{\mathrm{G}})\right| = \left|\mathcal{F}_y'(\mathcal{J}_1(x_{\mathrm{G}}^{(k)}))\right| = \left|\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) - \mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)}))\right|$$

  can be controlled by $c_0\tau$.
- The computation of the fixed point iteration in Algorithm 2 cannot get stuck in practical applications because the iteration will stop when the denominator of $\mathcal{J}_2(x)$ (i.e., $|\mathcal{J}_1(\mathcal{J}_1(x_{\mathrm{G}}^{(k)})) - 2\mathcal{J}_1(x_{\mathrm{G}}^{(k)}) + x_{\mathrm{G}}^{(k)}|$) is smaller than the given tolerance τ.

Putting everything together, Algorithm 1 gives the formal description of the proposed method for solving the original non-convex optimization problem (3). Note that Algorithm 1 calls GAI (Algorithm 2) as a subroutine.

Having fully described the algorithm, there are still several important questions remaining. Does the GAI algorithm converge at all? If so, at what speed? Can GAI obtain a global optimal solution (rather than just a stationary point) to (4)? It turns out that positive answers can be obtained for all of these questions. We will address them in detail next.

3.3 Convergence Analysis of the Fixed Point Algorithm

We are now ready to characterize the convergence speed of Algorithm 2, which is the computational bottleneck in Algorithm 1 (note that all other steps in Algorithm 1 take only constant time). Further, since Algorithm 2 only takes constant time if $\mathcal{F}_y'(a_0) >= 0$, it suffices to look at the non-trivial case where neither is true. The next theorem formalizes the result.

Theorem 1. *Given* $\mathcal{G}$ *satisfies Assumptions* ***A1-A3****. If* $\mathscr{S} \neq \emptyset$, *and* $\mathcal{F}_y'(a_0) < 0$, *then:*

i) *The sequence* $\{x_{\mathrm{L}}^{(k)}\}$ *generated by* $x_{\mathrm{L}}^{(k+1)} = \mathcal{J}_1(x_{\mathrm{L}}^{(k)})$ *with the initialization* $x_{\mathrm{L}}^{(0)} = y$ *converges to* $\bar{x}_y$ *at an (asymptotic) linear rate;*
ii) *The sequence* $\{x_{\mathrm{G}}^{(k)}\}$ *generated in Algorithm 1, i.e.,* $x_{\mathrm{G}}^{(k+1)} = \mathcal{J}_2(x_{\mathrm{G}}^{(k)})$ *with the initialization* $x_{\mathrm{G}}^{(0)} = y$ *converges to* $\bar{x}_y$ *at an (asymptotic) super-linear rate;*
iii) *For any* $k > 0$, $|x_{\mathrm{G}}^{(k)} - \bar{x}_y| = O(\rho^k)$, *for some* $0 < \rho < 1$.

The proof of Theorem 1 is given in Appendix, available in the online supplemental material. The key for the proof of Theorem 1 is to check the following three crucial convergence properties in turn:

1) $\lim_{x \to \bar{x}_y^+} \mathcal{J}_2(x) = \bar{x}_y$;
2) $\exists\, 0 < \rho < 1$ such that $|\mathcal{J}_2(x) - \bar{x}_y| < \rho|x - \bar{x}_y|$;
3) $\lim_{x \to \bar{x}_y^+} \mathcal{J}_2'(x) = 0$.

In Theorem 1, we prove the proposed iterative scheme (GAI) has a global contractive property and converges at an (asymptotic) super-linear rate to a globally optimal solution of (4), which is even faster than Lu's work.

3.4 Proximal Gradient Algorithm for a Generalized Problem

In this subsection, a more general framework is considered as follows.

$$\min_X \mathcal{D}(X) = \mathcal{L}(X) + \lambda \sum_{i=1}^{m} \mathcal{G}(\|\mathbf{x}_i\|_2), \tag{19}$$

where the function $\mathcal{G}: [0, +\infty) \to [0, +\infty)$ satisfies the assumptions **A1-A3**, and the function $\mathcal{L}: \mathbb{R}^{m \times n} \to [0, +\infty)$ has a Lipschitz continuous gradient with the Lipschitz constant being denoted as $l(\mathcal{L})$, i.e., $\|\nabla\mathcal{L}(A) - \nabla\mathcal{L}(B)\|_F \leq l(\mathcal{L})\|A - B\|_F$ holds for all $A, B \in \mathbb{R}^{m \times n}$. The proximal gradient (PG) algorithm [44] solves the problem (19) by the following updating rule [6]

6. Proximal gradient algorithm is wildly used in the following optimization problem: $\arg\min_X \mathcal{L}(X) + \mathcal{W}(X)$, in which $\mathcal{L}(\cdot)$ is convex and differentiable with a Lipschitz continuous gradient, and $\mathcal{W}(\cdot)$ is a convex and lower semicontinuous function. For this optimization problem, we can update X by $X^{(k)} = \mathrm{prox}_{\frac{1}{\xi}\mathcal{W}(\cdot)}(X^{(k-1)} - \frac{1}{\xi}\nabla\mathcal{L}(X^{(k-1)}))$, in which $\mathrm{prox}_{\frac{1}{\xi}\mathcal{W}(\cdot)}(O) = \arg\min_X \frac{1}{2}\|O - X\|_F^2 + \frac{1}{\xi}\mathcal{W}(X)$, $\xi > l(\mathcal{L})$, and $l(\mathcal{L})$ is a Lipschitz constant of the gradient of $\nabla\mathcal{L}$.

$$\begin{aligned} X^{(k+1)} &= \arg\min_X \mathcal{L}(X^{(k)}) + \left\langle \nabla\mathcal{L}(X^{(k)}), X - X^{(k)} \right\rangle \\ &\quad + \frac{\xi}{2}\|X - X^{(k)}\|_F^2 + \lambda\sum_{i=1}^{m}\mathcal{G}(\|\mathbf{x}_i\|_2) \\ &= \arg\min_X \frac{1}{2}\left\|X - O^{(k)}\right\|_F^2 + \frac{\lambda}{\xi}\sum_{i=1}^{m}\mathcal{G}(\|\mathbf{x}_i\|_2) \\ &= \mathcal{S}(O^{(k)}, \frac{\lambda}{\xi}), \end{aligned} \tag{20}$$

where $O^{(k)} = X^{(k)} - \frac{\nabla\mathcal{L}(X^{(k)})}{\xi}$, and $\xi > l(\mathcal{L})$. The convergence property for the iterate (20) can be guaranteed by the following theorem.

Theorem 2. *If $\xi > l(\mathcal{L})$, the sequence $\{X^{(k)}\}$ generated by (20) has the following properties.*

i) The objective value sequence $\{\mathcal{D}(X^{(k)})\}$ is monotonically decreasing. Specifically,

$$\mathcal{D}(X^{(k)}) - \mathcal{D}(X^{(k+1)}) \geq \frac{\xi - l(\mathcal{L})}{2}\left\|X^{(k+1)} - X^{(k)}\right\|_F^2.$$

ii) $\lim_{k\to+\infty}\|X^{(k)} - X^{(k+1)}\|_F = 0$.

iii) Suppose $\lim_{\|X\|_F\to+\infty}\mathcal{D}(X) = +\infty$, then any limit point of $\{X^{(k)}\}$ is a stationary point.

The proof of Theorem 2 is given in Appendix, available in the online supplemental material.

4 Three Applications of Structured Sparsity

Structured sparsity has been widely applied in many fields due to its good performance in studying the structured relationships hidden in samples, including outlier pursuit[45], [46], [47], feature selection[48], [49], dictionary learning[41], [50], background modeling and object detecting[51], [52], [53], [54], [55], [56], robust orthonormal subspace learning [57], effective matrix recovery [58], and low rank representation [59], [60]. In this section, we study three concrete problems (including outlier pursuit, supervised feature selection, and dictionary learning) that have wide-spread applications in computer vision, where structured sparsity can be leveraged to achieve good recovery results. In each of the three problems, we first describe the problem formulation and then how the problem can be relaxed by using surrogate functions. Subsequently, we describe the optimization algorithm for each problem, and show that the key computation step boils down to solving the structured sparsity optimization problem (3), to which Algorithm 1 readily applies. In addition, we provide theoretical guarantees for the overall optimization algorithms.

4.1 Outlier Pursuit

In many visual tasks, samples in multiple classes approximately lie in multiple low-dimensional subspaces [61], [62]. Thus, a matrix $M = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n]$ in which each column is a data sample has a low-rank structure. However, in real applications, observations are often grossly corrupted or even suffer from outliers, breaking the low-rank structure of the data. With the assumption that a small fraction of observations are outliers (or grossly corrupted data), the objective for locating the outliers (or recovering the low-rank data) can be formulated as follows [45].

$$\begin{aligned} (\text{OP}) \min_{X,E}\ & \text{rank}(X) + \lambda\|E\|_{2,0} \\ \text{s.t. } & X + E = M, \end{aligned} \tag{21}$$

where the matrix M of observations with outliers (or corrupted data) can be decomposed into a low-rank component X with no outliers (or noise) and a structured-sparsity component E with outliers (or noise). Here, $\ell_{2,0}$-norm regularization is imposed on E to characterize the small number of outliers (or grossly corrupted samples). The Outliers Pursuit (OP) task is the focus of this subsection. It is worth noting that finding the locations of non-zero columns of E is preferred over exactly recovering E in the OP task. This is because it is difficult to exactly recover E by optimizing problem (21) [47]. Determining the locations for the OP task is adequate since the non-zero columns of E correspond to the outliers in M.

The optimization problem (21) is not directly tractable because minimization of the functions of matrix rank and $\ell_{2,0}$-norm is NP-hard. To make problem (21) solvable, we replace the ℓ_0-norm with non-convex surrogate functions $\mathcal{G}$, and thus yields a relaxed problem

$$\begin{aligned} \min_{X,E}\ & \sum_{i=1}^{r}\mathcal{G}(\sigma_i(X)) + \lambda\sum_{i=1}^{n}\mathcal{G}(\|\mathbf{e}_i\|_2) \\ \text{s.t. } & X + E = M, \end{aligned} \tag{22}$$

where $r = \text{rank}(X)$. The equality constraint in problem (22) is replaced with a penalty function $\frac{1}{2}\|M - X - E\|_F^2$, and the following unconstrained optimization is solved instead.

$$\begin{aligned} \min_{X,E}\mathcal{Q}(X,E) = \alpha\sum_{i=1}^{r}\mathcal{G}(\sigma_i(X)) + \beta\sum_{i=1}^{n}\mathcal{G}(\|\mathbf{e}_i\|_2) \\ + \frac{1}{2}\|M - X - E\|_F^2. \end{aligned} \tag{23}$$

Optimizing X and E simultaneously in problem (23) could be expensive in practice. It is solved iteratively in this work by combining the coordinate descent algorithm with the proximal algorithm[44], as detailed below.

Step1 Given $X^{(k)}$ and $E^{(k)}$, update X by

$$\begin{aligned} X^{(k+1)} &= \arg\min_X \alpha\sum_{i=1}^{r}\mathcal{G}(\sigma_i(X)) + \eta\left\|X - X^{(k)}\right\|_F^2 \\ &\quad + \frac{1}{2}\left\|M - X - E^{(k)}\right\|_F^2 \\ &= \arg\min_X \frac{\alpha}{1+2\eta}\sum_{i=1}^{r}\mathcal{G}(\sigma_i(X)) + \frac{1}{2}\left\|X - O^{(k)}\right\|_F^2 \\ &= \mathcal{T}\left(O^{(k)}, \frac{\alpha}{1+2\eta}\right), \end{aligned} \tag{24}$$

where $O^{(k)} = \frac{M - E^{(k)} + 2\eta X^{(k)}}{1+2\eta}$ and $\eta > 0$.

Step2 Given $X^{(k+1)}$ and $E^{(k)}$, update E by

$$\begin{aligned} E^{(k+1)} &= \arg\min_{E} \beta \sum_{i=1}^{n} \mathcal{G}(\|\mathbf{e}_i\|_2) + \eta \left\|E - E^{(k)}\right\|_F^2 \\ &\quad + \frac{1}{2}\left\|M - X^{(k+1)} - E\right\|_F^2 \\ &= \arg\min_{E} \frac{\beta}{1+2\eta} \sum_{i=1}^{n} \mathcal{G}(\|\mathbf{e}_i\|_2) + \frac{1}{2}\left\|E - Q^{(k)}\right\|_F^2 \\ &= \mathcal{S}\Big(Q^{(k)}, \frac{\beta}{1+2\eta}\Big), \end{aligned} \tag{25}$$

where $Q^{(k)} = \frac{M - X^{(k+1)} + 2\eta E^{(k)}}{1+2\eta}$.

The convergence property of the above alternate iterate can be guaranteed by the following theorem.

Theorem 3. *For $\eta > 0$, the sequence $\{(X^{(k)}, E^{(k)})\}$ generated by alternative iterate of (24) and (25) satisfies the following properties.*

i) The objective value sequence $\{\mathcal{Q}(X^{(k)}, E^{(k)})\}$ is monotonically decreasing. Specifically,

$$\begin{aligned} &\mathcal{Q}(X^{(k)}, E^{(k)}) - \mathcal{Q}(X^{(k+1)}, E^{(k+1)}) \\ &\geq \eta(\|X^{(k+1)} - X^{(k)}\|_F^2 + \|E^{(k+1)} - E^{(k)}\|_F^2). \end{aligned} \tag{26}$$

ii) $\lim_{k\to+\infty} \|[X^{(k)}, E^{(k)}] - [X^{(k+1)}, E^{(k+1)}]\|_F = 0.$

The proof of Theorem 3 is given in Appendix, available in the online supplemental material.

In outlier pursuit, [46] has proposed a dictionary-based outlier pursuit method that can be regarded as a generalization of (21):

$$\begin{aligned} \min_{X,C} \quad & \operatorname{rank}(X) + \lambda\|C\|_{2,0} \\ \text{s.t.} \quad & \|M - X - DC\|_F \leq \epsilon_{\text{noise}}, \end{aligned} \tag{27}$$

where D is a known dictionary, ϵ_{noise} is error and the non-zero columns of DC are outliers not in the column space of X. Similar to the idea discussed in the above (optimization procedure for the non-convex version of (21)), we can solve the non-convex version of (27) instead, as below.

$$\begin{aligned} \min_{X,C} \quad & \sum_{i=1}^{r} \mathcal{G}(\sigma_i(X)) + \lambda \sum_{i=1}^{n} \mathcal{G}(\|\mathbf{c}_i\|_2) \\ \text{s.t.} \quad & \|M - X - DC\|_F \leq \epsilon_{\text{noise}}, \end{aligned} \tag{28}$$

By introducing a penalty function $\frac{1}{2}\|M - X - DC\|_F^2$, (28) becomes

$$\begin{aligned} \min_{X,C} \mathcal{Q}_{\text{D}}(X, C) &= \alpha \sum_{i=1}^{r} \mathcal{G}(\sigma_i(X)) + \beta \sum_{i=1}^{n} \mathcal{G}(\|\mathbf{c}_i\|_2) \\ &\quad + \frac{1}{2}\|M - X - DC\|_F^2. \end{aligned} \tag{29}$$

The optimization procedure for solving (29) can be described as following two steps:

Step1 Given $X^{(k)}$ and $C^{(k)}$, update X by

$$\begin{aligned} X^{(k+1)} &= \arg\min_{X} \alpha \sum_{i=1}^{r} \mathcal{G}(\sigma_i(X)) + \eta \left\|X - X^{(k)}\right\|_F^2 \\ &\quad + \frac{1}{2}\left\|M - X - C^{(k)}\right\|_F^2 \\ &= \arg\min_{X} \frac{\alpha}{1+2\eta} \sum_{i=1}^{r} \mathcal{G}(\sigma_i(X)) + \frac{1}{2}\left\|X - O_{\text{D}}^{(k)}\right\|_F^2 \\ &= \mathcal{T}\Big(O_{\text{D}}^{(k)}, \frac{\alpha}{1+2\eta}\Big), \end{aligned} \tag{30}$$

where $O_{\text{D}}^{(k)} = \frac{M - E^{(k)} + 2\eta X^{(k)}}{1+2\eta}$ and $\eta > 0$.

Step2 Given $X^{(k+1)}$ and $C^{(k)}$, update C by

$$\begin{aligned} C^{(k+1)} &= \arg\min_{E} \beta \sum_{i=1}^{n} \mathcal{G}(\|\mathbf{c}_i\|_2) + \eta \left\|C - C^{(k)}\right\|_F^2 \\ &\quad + \frac{1}{2}\left\|M - X^{(k+1)} - DC\right\|_F^2 \end{aligned} \tag{31}$$

where (31) can be solved by the proposed proximal gradient algorithm in Section 3.4.

4.2 Supervised Feature Selection

Feature selection (FS) is a natural application of sparse representation theory that seeks sparse and representative features from input data. Structured sparsity based FS methods [48], [49] have been gaining attention recently. When compared to traditional FS methods [63], [64], [65], [66], the joint evaluation mechanism makes it more efficient to extract representative and discriminative features and more robust to data noise. Given a dataset $\{(\mathbf{m}_i, \mathbf{l}_i)\}_{i=1}^{n}$ with ι classes, where $\mathbf{m}_i \in \mathbb{R}^d$ is the i-th training sample and $\mathbf{l}_i \in \mathbb{R}^{\iota}$ is the corresponding one-hot label vector (i.e., if $\mathbf{m}_i$ is from the j-th class, then the j-th entry of $\mathbf{l}_i$ is 1, and the rest of the entries are 0). Let $M = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n]$, and $L = [\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n]$, then the classical structured sparsity based FS model can be formulated as

$$\min_{W} \|WM - L\|_F^2 + \lambda\|W\|_{2,0}, \tag{32}$$

where $W \in \mathbb{R}^{\iota \times d}$ is a projection matrix, and $\|W\|_{2,0}$ is used to select the most representative and discriminative features across all data samples with joint sparsity. Similarly, the $\ell_{2,0}$-norm regularized problem (32) is NP-hard. To make it solvable, the $\ell_{2,0}$-norm is replaced with some suitable non-convex surrogate functions $\mathcal{G}$, and (32) is relaxed as

$$\min_{W} \frac{1}{2}\|WM - L\|_F^2 + \lambda \sum_{i=1}^{d} \mathcal{G}(\|\mathbf{w}_i\|_2), \tag{33}$$

where $\mathbf{w}_i$ is the i-th column of the matrix W.

Next, we discuss how to solve this problem. To remove the interdependence of W in the two terms of the objective, an auxiliary matrix Q is introduced and the problem (33) becomes

$$\begin{aligned} \min_{Q,W} \quad & \frac{1}{2}\|WM - L\|_F^2 + \lambda \sum_{i=1}^{d} \mathcal{G}(\|\mathbf{q}_i\|_2) \\ \text{s.t.} \quad & W = Q, \end{aligned} \tag{34}$$

where $\mathbf{q}_i$ is the i-th column of Q. The problem (34) can be efficiently solved by the framework of alternating direction method of multipliers (ADMM). To do so, we write out the augmented Lagrangian function for problem (34) as follows.

$$\mathcal{L}_\mu(W,Q,\Lambda) = \frac{1}{2}\|WM-L\|_F^2 + \lambda\sum_{i=1}^{d}\mathcal{G}(\|\mathbf{q}_i\|_2) + \langle\Lambda, W-Q\rangle + \frac{\mu}{2}\|W-Q\|_F^2, \tag{35}$$

where Λ is a Lagrange multiplier matrix and μ is a positive scalar. The whole algorithm proceeds as follows.

Step1 Given $Q^{(k)}$ and $\Lambda^{(k)}$, update W by

$$\begin{aligned} W^{(k+1)} &= \arg\min_W \mathcal{L}_\mu(W, Q^{(k)}, \Lambda^{(k)}) \\ &= \arg\min_W \frac{1}{2}\|WM-L\|_F^2 + \frac{\mu}{2}\left\|W - Q^{(k)} + \frac{\Lambda^{(k)}}{\mu}\right\|_F^2 \\ &= (\mu Q^{(k)} - \Lambda^{(k)} + LM^T)(\mu I + MM^T)^{-1}. \end{aligned} \tag{36}$$

Step2 Given $W^{(k+1)}$ and $\Lambda^{(k)}$, update Q by

$$\begin{aligned} Q^{(k+1)} &= \arg\min_Q \mathcal{L}_\mu(W^{(k+1)}, Q, \Lambda^{(k)}) \\ &= \arg\min_Q \frac{\lambda}{\mu}\sum_{i=1}^{m}\mathcal{G}(\|\mathbf{q}_i\|_2) + \frac{1}{2}\|O^{(k+1)} - Q\|_F^2 \\ &= \mathcal{S}(O^{(k+1)}, \frac{\lambda}{\mu}), \end{aligned} \tag{37}$$

where $O^{(k+1)} = W^{(k+1)} + \frac{1}{\mu}\Lambda^{(k)}$. It is easy to see that the problem (37) can be solved by Algorithm 1.

Step3 Given $W^{(k+1)}$, $Q^{(k+1)}$, and $\Lambda^{(k)}$, update the Lagrange multiplier matrix Λ by

$$\Lambda^{(k+1)} = \Lambda^{(k)} + \mu(W^{(k+1)} - Q^{(k+1)}).$$

Although global convergence of the ADMM can not be guaranteed because of the non-convexity of the objective, the ADMM has demonstrated superior performance in solving non-convex problems in practice [22], [23], [43], [67].

4.3 Structured Dictionary Learning

4.3.1 Structured Dictionary Learning

Structured sparsity has been successfully integrated into the dictionary learning framework in recent years [41], [50]. In this subsection, we focus on a general formulation of structured sparsity based dictionary learning with n samples in ι classes as follows.

$$\min_{D\in\mathscr{D},Z}\frac{1}{2}\|M-DZ\|_F^2 + \lambda\sum_{i=1}^{\iota}\sum_{j=1}^{m}\mathcal{G}(\|\mathbf{z}_i^j\|_2). \tag{38}$$

Here, $M=[M_1,M_2,\ldots,M_\iota]\in\mathbb{R}^{d\times n}$ is the matrix of data samples, where M_i is the data matrix of the i-th class; $D=[\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_m]\in\mathbb{R}^{d\times m}$ is the dictionary; $Z=[Z_1,Z_2,\ldots,Z_\iota]$ is the sparse coefficient matrix of data samples over the dictionary D, where Z_i is the coefficient matrix of data samples belonging to the i-th class; $\mathbf{z}_i^j$ is the j-th row vector of Z_i; $\mathscr{D}=\{D\in\mathbb{R}^{d\times m}, \mathbf{d}_j^\top\mathbf{d}_j\le 1, \forall\, j=1,2,\ldots,m\}$; $\|M-DZ\|_F^2$ is the reconstruction error term, and the non-convex surrogate term $\sum_{i=1}^{\iota}\sum_{j=1}^{m}\mathcal{G}(\|\mathbf{z}_i^j\|_2)$ is to effectively exploit the common features shared by the same class, and thus increase the discrimination and robustness of the learned dictionary.

The optimization for the problem (38) can be divided into two sub-problems: updating Z by fixing D, and updating D by fixing Z. When D is fixed, the sub-problem with respect to Z can be formulated as

$$\min_Z\sum_{i=1}^{\iota}\left\{\frac{1}{2}\|M_i-DZ_i\|_F^2 + \lambda\sum_{j=1}^{m}\mathcal{G}(\|\mathbf{z}_i^j\|_2)\right\}, \tag{39}$$

which can be separated into ι independent sub-problems ($i=1,2,\ldots,\iota$)

$$\min_{Z_i}\frac{1}{2}\|M_i-DZ_i\|_F^2 + \lambda\sum_{j=1}^{m}\mathcal{G}(\|\mathbf{z}_i^j\|_2) \tag{40}$$

because Z_i's in (39) are independent. To remove the interdependency of the terms in the objective (40), an auxiliary variable A_i is introduced and the the problem (40) becomes

$$\min_{Z_i,A_i}\frac{1}{2}\|M_i-DA_i\|_F^2 + \lambda\sum_{j=1}^{m}\mathcal{G}(\|\mathbf{z}_i^j\|_2), \quad \text{s.t. } Z_i=A_i. \tag{41}$$

When Z is fixed, the sub-problem with respect to D is

$$\min_{D\in\mathscr{D}}\frac{1}{2}\|M-DZ\|_F^2. \tag{42}$$

To remove the interdependency between the objective function and the constraint set $\mathscr{D}$ for variable D, we introduce an auxiliary matrix G and an indicator function

$$\mathcal{I}_{\mathscr{D}}(G) = \begin{cases} 0 & \text{if } G\in\mathscr{D}, \\ +\infty & \text{otherwise}, \end{cases} \tag{43}$$

and thus the problem (42) becomes

$$\min_{D,G}\frac{1}{2}\|M-DZ\|_F^2 + \mathcal{I}_{\mathscr{D}}(G), \quad \text{s.t. } D=G. \tag{44}$$

Both problems (41) and (44) can be solved by the framework of ADMM. The major difference is that the convergence for the former can not be guaranteed because of the non-convexity of the objective, while that for the latter can be guaranteed because the objective is convex and there are only two block variables [68].

When the dictionary D is obtained, the atom set of the i-th class is defined as

$$\mathscr{D}_i = \{\mathbf{d}_j |\ \|\mathbf{z}_i^j\|_2 > 0\}, \quad i=1,\ldots,\iota. \tag{45}$$

Accordingly, the i-th class-specific dictionary $D_i\in\mathbb{R}^{d\times|\mathscr{D}_i|}$ is constructed by using all $\mathbf{d}_j\in\mathscr{D}_i$ as its columns. Thus, classifying an unlabeled sample $\mathbf{m}_{\text{test}}$ is performed by the following three steps.

Step 1: Calculate the sparse representation for $\mathbf{m}_{\text{test}}$ over each class-specific dictionary D_i ($i=1,\ldots,\iota$) by

$$\hat{\mathbf{z}}_i = \arg\min_{\mathbf{z}}\|\mathbf{m}_{\text{test}} - D_i\mathbf{z}\|_2^2 + \lambda\|\mathbf{z}\|_1. \tag{46}$$

Step 2: Calculate the reconstruction error for $\mathbf{m}_{\text{test}}$ with respect to $D_i (i = 1, \ldots, \iota)$ by

$$e_i = \|\mathbf{m}_{\text{test}} - D_i \hat{\mathbf{z}}_i\|_2. \tag{47}$$

Step 3: Predict the class label of sample $\mathbf{m}_{\text{test}}$ by

$$l_{\text{test}} = \arg\min_i \{e_i\}. \tag{48}$$

4.3.2 Robust Structured Dictionary Learning

As the traditional dictionary learning methods are often sensitivity to outlier samples, [41] adopts $\sum_{k=1}^{n} \|\mathbf{m}_k - D\mathbf{z}_k\|_2^p$ instead of $\frac{1}{2}\|M - DZ\|_F^2$ in (38) to constrain reconstruction error in the dictionary learning model (i.e., $\mathcal{G}$ is taken as ℓ_p norm in $\sum_{j=1}^{m} \mathcal{G}(\|\mathbf{z}_i^j\|_2)$), where $\mathbf{m}_k$ and $\mathbf{z}_k$ are used to stand for the k-th column vector in M and Z. Therefore, we consider a more complex case of structured dictionary learning, i.e., using $\sum_{k=1}^{n} \mathcal{G}(\|\mathbf{m}_k - D\mathbf{z}_k\|_2)$ in (38) instead of $\frac{1}{2}\|M - DZ\|_F^2$,

$$\min_{D \in \mathscr{D}, Z} \sum_{k=1}^{n} \mathcal{G}(\|\mathbf{m}_k - D\mathbf{z}_k\|_2) + \lambda \sum_{i=1}^{\iota} \sum_{j=1}^{m} \mathcal{G}(\|\mathbf{z}_i^j\|_2). \tag{49}$$

To solve (49) effectively, two auxiliary variables, H and G, are introduced and the problem (49) becomes

$$\begin{aligned} &\min_{D,Z} \sum_{k=1}^{n} \mathcal{G}(\|\mathbf{h}_i\|_2) + \lambda \sum_{i=1}^{\iota} \sum_{j=1}^{m} \mathcal{G}(\|\mathbf{z}_i^j\|_2) + \mathcal{I}_{\mathscr{D}}(G) \\ &\text{s.t. } H = M - DZ, \quad D = G, \end{aligned} \tag{50}$$

which can be solved by the framework of ADMM. When the dictionary D is obtained, we can classify the unlabeled sample $\mathbf{m}_{\text{test}}$ by the Steps 1-3 in the Section 4.3.1.

5 Experimental Results

In this section, a variety of simulations and experimental evaluations are presented to demonstrate the efficiency and effectiveness of the proposed algorithm in the structured sparsity framework. We divide the presentation into four subsections.

In the first subsection, we provide simulation results that verify the convergence rate of GAI (Algorithm 2). Additionally, we perform some comparisons between the GAI and Lu's work [27] described in Algorithm 4, which was also designed to solve the problem (4) but for a different purpose. As made clear in the simulation results, the proposed algorithm converges to a global optimal solution significantly faster.

The last three subsections present three real-world applications in which structured sparsity can be harnessed to achieve good recovery performance. These three applications are outlier pursuit, feature selection, and structured dictionary learning. Detailed discussions of their formulations and optimization algorithms have been provided in the previous section. Note that in each of the three applications, one can select many surrogate functions since our framework allows for a general set of such surrogate functions, and we have kept discussions of optimization algorithms generic in the previous section. Here, we will instantiate a wide variety of surrogate functions and compare their performance. See below for a quick list.

- ℓ_1: Algorithm 1 is combined with ℓ_1-norm soft threshold to solve the problem (2).
- Structural Sparse Preserving via Mixed $\ell_{2,p}$ Norm (SSP-ℓ_p): Proposition 2 in [24] is combined with generalized soft-thresholding (GST) [29] to solve the problem (3) with non-convex surrogate ℓ_p-norm.
- The methods combining Algorithm 1 with GAI for non-convex surrogate functions ℓ_p-norm, Geman penalty, Laplace penalty, LOG penalty, Logarithm penalty, and ETP penalty are denoted as ℓ_p, Geman, Laplace, LOG, Logarithm, and ETP, respectively.

5.1 Convergence Speed of the Fixed Point Algorithm

Here we empirically evaluate the convergence speed and give some comparisons between the proposed fixed point algorithm and Lu's work. Lu's work is given in Algorithm 4. The following problem is considered.

$$\arg\min_X \mathcal{Z}(X) = \frac{1}{2}\|Y - X\|_F^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{n} \mathcal{G}(|x_{ij}|), \tag{51}$$

where $X, Y \in \mathbb{R}^{m \times n}$, and x_{ij} is the element of X.

In the simulations, the entries of the matrix $Y \in \mathbb{R}^{100 \times 100}$ are i.i.d. from the standard normal distribution $\mathcal{N}(0, 1)$, and the value of parameter λ is set to 1. For each non-convex surrogate function, the output of Lu's work with $\kappa = 1000$ and $\tau = 10^{-20}$ is taken as the ground truth, which is denoted as X^*. *Absolute Error* $= |\mathcal{Z}(X^*) - \mathcal{Z}(\hat{X}_k)|$ is used to evaluate the performance of Lu's work and the proposed fixed point algorithm for solving the problem (51), where $\hat{X}_k$ is the output of the k-th iteration. Obviously, for a fixed number k, a smaller *Absolute Error* indicates a faster convergence rate.

The simulation results in Fig. 1 provide a clear demonstration of the effectiveness of the proposed fixed point algorithm. The curves of the proposed algorithm show a sharp decrease during the iteration process, and the corresponding *Absolute Errors* reach 10^{-7} within 6 iterations at most. In contrast, the curves of Lu's work decrease slowly, and the number of iterations is even more than 100 for the surrogate function Laplace when the corresponding *Absolute Errors* reaches 10^{-7}. All these results validate the theoretical characterizations and demonstrate the superiority of the proposed fixed point algorithm.

5.2 Outlier Pursuit

In this part, we apply the proposed framework to the OP task on the handwritten digit dataset `MNIST` and the point trajectory dataset `Hopkins 155`. To give numerical comparisons, two metrics, including *Ham* and *F1* [69], are chosen to evaluate the location differences of non-zero columns between the ground truth E_0 (corresponding outliers within the data) and the recovery E^* in (21). The *Ham* metric is formulated as $Ham = \frac{100-h}{100}$, where h is the Hamming distance [47] between the column supports of E_0 and E^*. The *F1* metric is formulated as $F1 = \frac{2 \times P \times R}{P+R}$, where $P = \frac{TP}{TP+FN}$, $R = \frac{TN}{TN+FP}$, and TP, FP, TN, and FN are true positive, false positive, true negative, and false negative for whether one can

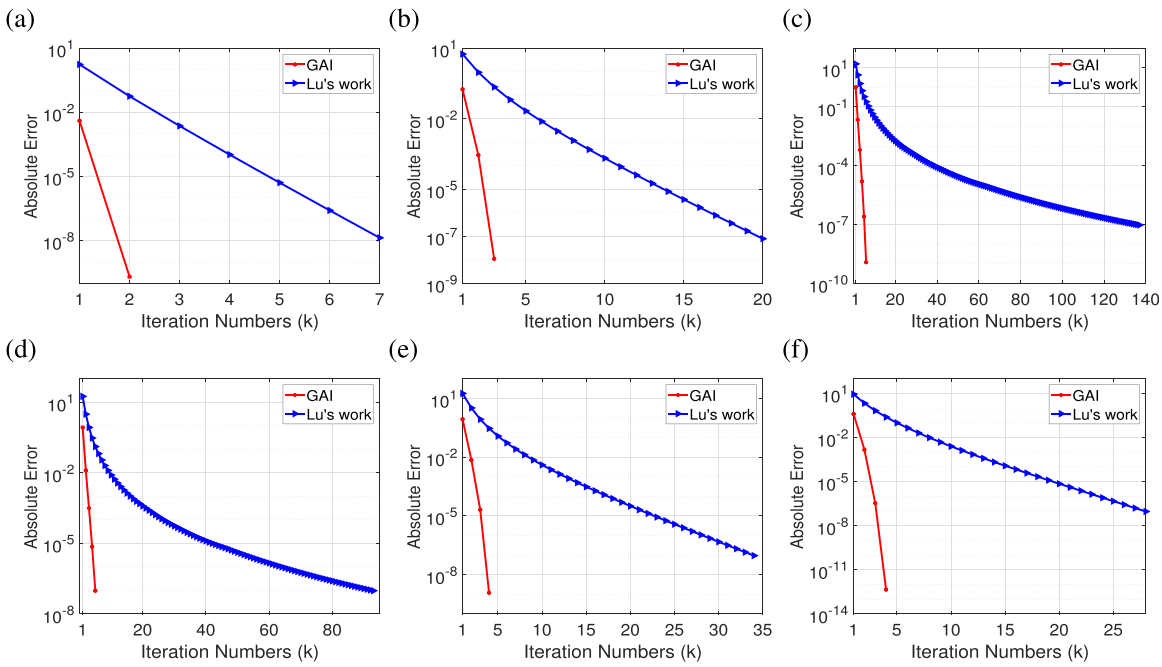


Fig. 1. Comparisons between the proposed algorithm GAI and Lu's work with the surrogate $\mathcal{G}$ being (a) ℓ_p norm, $p = 0.5$, (b) Geman penalty, $\gamma = 1$, (c) Laplace penalty, $\gamma = 1$, (d) LOG penalty, $\gamma = 1$, (e) Logarithm penalty, $\gamma = 1$, and (f) ETP, $\gamma = 1$.

correctly find the locations of non-zero columns, respectively. For each dataset, experiments are repeated 20 times with randomly chosen outliers, and the average values of *Ham* and *F1* are reported. Three baseline methods, including Robust Principal Component Analysis (RPCA) [13], Robust Principal Component Analysis via Outlier Pursuit (RPCA-OP) [45], [7] and SSP-ℓ_p, are used to compare to (22) with different surrogate functions.

5.2.1 Results on the `MNIST` Dataset

From digit classes '1'-'9', c_0 images are randomly chosen as outliers; from digit class '0', the first $100 - c_0$ images are chosen as the intrinsic samples. The observation matrix D in problem (23) is constructed by these chosen samples. The goal of this experiment is to compare the capability of different surrogates of $\ell_{2,0}$-norm for locating the position of outliers.

The values of *Ham* and *F1* for different numbers c_0 of outliers are recorded in Tables 3 and 4, respectively. It can be observed that the accuracies of outlier detection decrease as the number c_0 increases. Additionally, most of the non-convex surrogates (especially for Logarithm penalty) are more effective than the convex surrogate $\ell_{2,1}$-norm (RPCA-OP). This also verifies that non-convex surrogates have a superior approximation of the $\ell_{2,0}$-norm and thus detect outliers more precisely than $\ell_{2,1}$-norm. Compared with all baseline methods, ℓ_p and Logarithm outperform the baseline methods in most cases, where Logarithm achieves the best performance across all cases. These results illustrate the effectiveness of the proposed method. To provide a clear intuition of the outlier pursuit, the recovered low-rank part and the column-sparse noise part of identified outliers with $c_0 = 5$ by different surrogates are shown in Fig. 2. Due to the space limit, we only list the results of five methods, including RPCA, RPCA-OP, SSP-ℓ_p, ℓ_p and Logarithm (in which ℓ_p and Logarithm are chosen because of their good performance in both metrics of *Ham* and *F1*). As illustrated in Fig. 2, ℓ_p and Logarithm both find more outliers than ℓ_1 minimization does. The (22) based methods obtain the clearest recovered outliers.

5.2.2 Results on the `Hopkins 155` Dataset

Experiments are performed on three video sequences from the `Hopkins 155` dataset, including "1R2RC", "1R2RCR", and "three-cars". "1R2RC" is the sequence with two objects rotating for a fixed camera; "1R2RCR" is the sequence with all of the two objects and the camera rotating; "three-cars" contains three motions of two toy cars and a box moving on a table, and the motions are taken by a fixed camera. Some examples are illustrated in Fig. 3. As stated in [72], if the point trajectories associated with multiple moving objects lie in multiple low-dimensional subspaces, then the matrix constructed by the point trajectories has a low-rank structure. For each of the three video sequences, c_1 point trajectories with m_0-dimensional features in the first object are randomly

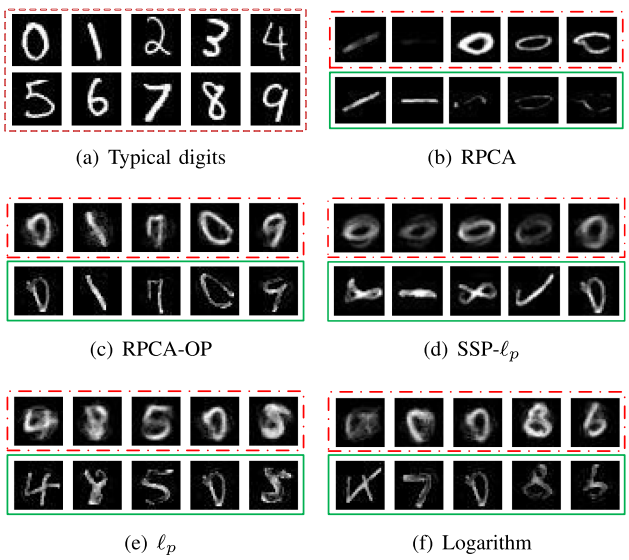
TABLE 3
Comparison of *Ham* for Anomaly Identification on the `MNIST` Dataset

Video Clip	c=5	c=10	c=15	c=20
RPCA [13]	0.9660	0.9080	0.8640	0.8400
RPCA-OP [45]	0.9680	0.9460	0.9060	0.8800
SSP-ℓ_p [24]	**0.9780**	0.9460	0.9160	0.8940
ℓ_p	0.9760	0.9480	**0.9220**	0.8840
Geman	0.9080	0.8200	0.7420	0.6660
Laplace	0.9680	0.9480	0.8960	0.8620
LOG	0.9100	0.8240	0.7520	0.6720
Logarithm	**0.9780**	**0.9560**	0.9180	**0.8980**
ETP	0.9480	0.9020	0.8560	0.7960

7. RPCA-OP can be regarded as the case of taking $\mathcal{G} = |\cdot|$ in (22).

TABLE 4
Comparison of *F1* Metric for Anomaly Identification on the MNIST Dataset

Video Clip	c=5	c=10	c=15	c=20
RPCA [13]	0.7000	0.5800	0.5600	0.6200
RPCA-OP [45]	0.7600	0.7200	0.7133	0.7150
SSP-ℓ_p [24]	0.7800	0.7200	0.7133	0.7050
ℓ_p	**0.8000**	0.7300	0.7200	0.7250
Geman	0.0800	0.1200	0.1467	0.1750
Laplace	0.7800	0.6900	0.6800	0.6600
LOG	0.1000	0.1300	0.1533	0.1850
Logarithm	0.7800	**0.7800**	**0.7267**	**0.7450**
ETP	0.4200	0.5400	0.4800	0.5150



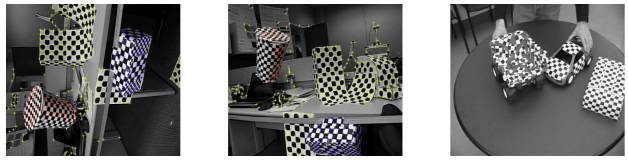
(a) Typical digits (b) RPCA (c) RPCA-OP (d) SSP-ℓ_p (e) ℓ_p (f) Logarithm

Fig. 2. Identified outliers by different surrogates. The subfigures surrounded by red lines and green lines are the visual representation of the low rank part X^* and sparse part E^*, respectively.

chosen as the intrinsic samples, and c_2 point trajectories from the second object are randomly chosen as outliers. These chosen point trajectories are used as columns of the observation matrix D in problem (23). In the experiments, (c_1, c_2, m_0) are set as (89,15,59), (50,35,49), and (30,30,31) for video sequences "1R2RC", "1R2RCR", and "three-cars", respectively.

The experimental results are presented in Fig. 4. It can be seen that most non-convex surrogates outperform the ℓ_1-norm (RPCA-OP). ETP even achieves perfect performance for both metrics of *Ham* and *F1* on the sequences "1R2RC" and "1R2RCR". In the case of the baseline methods versus the proposed methods (including ℓ_p, Logarithm, and ETP), the proposed methods outperform the baseline methods in most cases. In addition, to further illustrate the efficiency of the proposed algorithm, the running time is shown in Fig. 4c, from which it can be concluded that the time for solving the problem (22) with non-convex surrogates (except for ℓ_p, Geman and Laplace) is comparable to (or even less than) that with the ℓ_1-norm (RPCA-OP). In terms of running times, most of the methods based on our algorithm achieve better results than the baseline methods SSP-ℓ_p and RPCA.

The above experimental results illustrate the effectiveness and efficiency of Algorithm 1 in the application of OP tasks.



(a) (b) (c)

Fig. 3. Examples in Hopkins 155: (a) 1R2RC; (b) 1R2RCR; (c) three-cars.

5.3 Supervised Feature Selection

In this subsection, experiments are performed on datasets ALLAML, Carcinomas, and GLIOMA to study the performance of the proposed framework on the feature selection problem (33). For each dataset, 50% of samples are randomly selected for training, and the rest are for testing. The performance of feature selection is evaluated by classification accuracy via k nearest neighbor (kNN), where k is set as 1 in the experiments. Here, four baseline methods, including kNN (k=1), Robust Feature Selection Based on $\ell_{2,1}$-Norms (RFS) [70], Top-k Supervise Feature Selection (TK-SFS) [71] and SSP-ℓ_p are used to compare to (33) with different surrogate functions.

The accuracies of different surrogates with 20 and 80 selected features are reported in Table 5. According to the results, it can be observed that the non-convex surrogates achieve significantly better performance than the ℓ_1-norm. And, regardless of whether the cases are top 20 or top 80 features, all (33) based methods outperform the three baseline methods (including 1NN, RFS, and TK-SFS). Specifically, ETP almost has the best performance when the number of selected features is 20. In the case of the top 80 features, both ℓ_p norm and Laplace perform the best in most cases. This provides additional evidence of the effectiveness of the proposed framework for solving feature selection models with non-convex surrogates.

5.4 Structured Dictionary Learning

5.4.1 *Structured Dictionary Learning*

In this part, experiments are conducted on eight image datasets, including USPS, Extended YaleB, ORL, PIE, UMIST, COIL-20, SB-Data, and TDT2 for the structured dictionary learning problem. An overall description of these datasets is provided in Table 6, in which τ and ν respectively denote the numbers of dictionary atoms and the training samples per class. There are three parameters, λ, γ and p, in the experiments. λ, γ, and p are selected from $\{0.00001, 0.00005, \cdots, 0.01\}$, $\{0.00001, 0.00005, \cdots, 1, 5, 10\}$, and $\{0.1, 0.2, \cdots, 0.8, 0.9\}$, respectively. The experiments are repeated 10 times with different random splits of the datasets, and the average classification accuracies with the best parameters are reported. The classification accuracies on the eight benchmark datasets are recorded in Table 7. As demonstrated by the results, the best classification accuracy is distributed in non-convex surrogate functions for each dataset, confirming their superior performance.

Additionally, the nonparametric post-hoc statistic test [73], [74] is carried out to further explore the comprehensive capability of these non-convex surrogate functions. Here, the critical difference (CD) diagram is used to

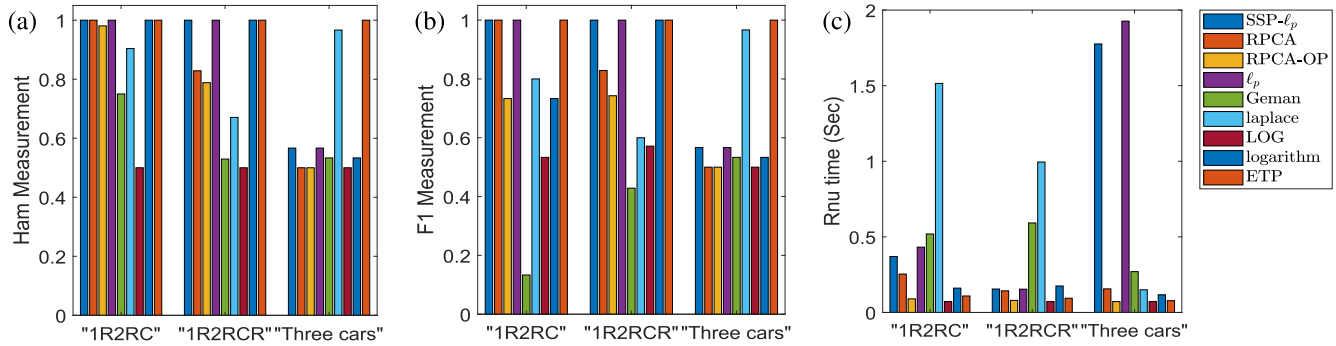


Fig. 4. Comparisons of (a) *Ham*, (b) *F1* metric, and (c) Running time (seconds) for anomalies identification on the `Hopkins 155` dataset.

TABLE 5
Classification Accuracy on the Task of Feature Selection

	Feature selection of top 20 features				Feature selection of top 80 features			
	ALLAML	Carcinomas	GLIOMA	Average	ALLAML	Carcinomas	GLIOMA	Average
1NN [69]	83.78	85.39	50.00	73.06	83.78	85.39	50.00	73.06
RFS [70]	83.78	85.39	50.00	73.06	83.78	85.39	50.00	73.06
TK-SFS[71]	67.57	77.53	26.92	57.34	67.57	71.91	50.00	63.16
SSP-ℓ_p[24]	91.89	**91.01**	73.08	85.33	94.59	95.51	**76.92**	**89.01**
ℓ_1	89.19	88.76	53.85	77.27	89.19	93.26	61.54	81.33
ℓ_p	91.89	**91.01**	73.08	85.33	94.59	95.51	**76.92**	**89.01**
Geman	91.89	88.76	69.23	83.29	97.30	92.13	73.08	87.50
Laplace	91.89	**91.01**	57.69	80.20	**100.00**	**96.63**	69.23	88.62
LOG	91.89	88.76	53.85	78.17	91.89	89.89	53.85	78.54
Logarithm	91.89	**91.01**	57.69	80.20	91.89	94.38	61.54	82.60
ETP	**97.30**	89.89	**88.46**	**91.88**	97.30	95.51	73.08	88.63

TABLE 6
Overall Description of the Datasets

Datasets	DIM	Data#	Class#	τ	ν
ORL	1024	400	40	80	7
PIE	1024	2040	12	240	80
TDT2	500	1560	30	210	31
USPS	256	1100	10	200	60
UMIST	750	575	20	100	15
COIL-20	1521	1440	20	200	40
SBData	638	2000	40	200	20
E-YaleB	1024	2414	38	380	20

TABLE 7
Classification Accuracy of Different Datasets

Datasets	ORL	COIL	PIE	TDT2	USPS	YaleB	UMIST	SBD
SSP-ℓ_p[24]	73.33	94.55	95.25	88.68	88.04	90.94	83.85	67.12
ℓ_1	76.43	94.71	95.28	88.35	87.87	**90.95**	83.83	67.11
ℓ_p	77.41	94.71	**95.27**	88.61	**88.18**	90.94	83.89	67.12
Geman	73.58	**95.69**	**95.27**	88.90	88.02	90.93	83.85	67.03
Laplace	**77.65**	94.23	95.24	87.01	86.51	90.77	**85.68**	66.92
LOG	20.43	94.75	95.24	**89.17**	87.50	90.94	83.85	**67.61**
Logarithm	76.54	94.62	95.32	88.38	87.95	90.94	83.83	67.11
ETP	77.56	93.15	94.92	87.05	86.52	88.94	85.52	67.59

intuitively illustrate the overall statistical comparisons of different non-convex surrogates on the eight datasets. The post-hoc statistical test results are graphically shown

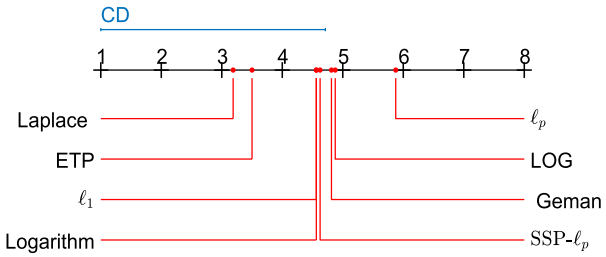


Fig. 5. Post-hoc statistical tests of different non-convex surrogate functions on overall datasets.

in Fig. 5, where the right side of the axis refers to the highest rank (best performance). As illustrated in Fig. 5, two non-convex surrogates of ETP and Laplace perform worse than the commonly used ℓ_1 surrogate, which demonstrates that ℓ_1 surrogate usually provides a stable performance for sparse representation. However, there are also four non-convex surrogate functions that achieve better performance than ℓ_1 norm, i.e., ℓ_p, LOG, Geman, and Logarithm. This is the main focus of this paper, which is an enlargement of the set of non-convex surrogate functions that are superior to ℓ_1 norm.

5.4.2 *Robust Structured Dictionary Learning*

In this part, we test four baseline methods, including Sparse Representation-based Classification (SRC)[6], K-SVD[75], Discriminative K-SVD (D-K-SVD)[76] and Semi-Supervised Robust Dictionary Learning (SSR-D) [41], and (49) with three non-convex surrogate functions (i.e., ℓ_p, Geman, and

TABLE 8
Classification Accuracy of Different Datasets

Datasets	AT&T	USPS	BinAlpha	TDT2
SRC[6]	97.00	93.91	74.68	92.87
K-SVD[75]	97.00	83.18	56.35	**96.33**
D-K-SVD[76]	97.00	84.18	56.11	96.13
SSR-D[41]	96.75	93.27	80.00	92.47
ℓ_p	**97.25**	95.27	**80.32**	91.31
Geman	97.00	93.45	80.00	91.27
Laplace	**97.25**	**95.36**	80.08	91.27

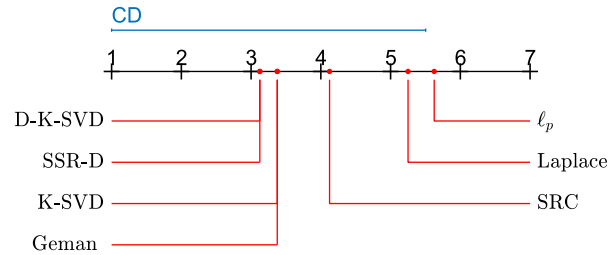


Fig. 6. Post-hoc statistical tests of all methods on overall datasets.

Laplace) on four datasets that are used in [41], including AT&T[8], USPS, BinAlpha[9], and TDT2).

We conduct standard five-fold cross-validation on each data set, and report the average classification accuracy by the seven methods on the four datasets in Table 8. As demonstrated by the results, the best classification accuracy is distributed in ℓ_p and Laplace for most datasets. Furthermore, as shown in Fig. 6, two non-convex surrogates of ℓ_p and Laplace provide stable performance for sparse representation when compared to other methods. These results also confirm the effectiveness of the proposed approach.

6 Conclusion

This paper provided a structured sparsity optimization framework that effectively harnesses structured sparsity in real-world problems. We moved beyond ℓ_1-norm based surrogate functions, and worked with a family of non-convex surrogate functions that are much more effective. We achieved this goal by exploring the relation between $\mathscr{A}$ and $\mathscr{B}$ in Lamma 1, which was also discussed in conference version [77] of this paper. But different from the conference version, in this paper, we further developed a novel iterative scheme (GAI) which solved the key sparsity optimization problem (4) to global optimality with geometric rate and helped to give a high-efficiency solver for sparsity and low rank recovery with non-convex surrogates. Based on GAI, two resulting algorithms, including structured sparsity optimization framework and GSVT-GAI, were proposed to solve the critical problems (3) and (6) for structured sparsity and matric low-rankness, respectively. Besides, to demonstrate the generality and wide applicability of the proposed algorithms, we presented three concrete problems (i.e., outlier pursuit, supervised feature selection, and structured dictionary learning), which can be solved by the proposed general structured sparsity optimization framework and GSVT-GAI effectively and efficiently. Extensive experiments on both synthetic data and real-world applications in the three concrete problems have validated the effectiveness and efficacy of the proposed framework and algorithms.

8. The AT&T dataset contains 400 data samples of 40 classes, and each data sample has 644 dimensions.

9. The BinAlpha dataset contains 1404 data samples of 36 classes, and each data sample has 320 dimensions.

References

[1] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.
[2] T. Geng, G. Sun, Y. Xu, and X. Liu, "Image compressed sensing recovery based on multi-scale group sparse representation," in *Proc. Int. Conf. Syst., Signals Image Process.*, 2018, pp. 1–5.
[3] M. Chen, F. Renna, and M. R. D. Rodrigues, "Compressive sensing with side information: How to optimally capture this extra information for GMM signals?," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2314–2329, May 2018.
[4] E. J. Candes and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Found. Comput. Math.*, vol. 6, no. 2, pp. 227–254, 2006.
[5] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2005.
[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[7] M. Harandi, C. Sanderson, C. Shen, and B. Lovell, "Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3120–3127.
[8] X. T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.
[9] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1 norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, 2006.
[10] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Commun. Pure Appl. Math.*, vol. 98, no. 6, pp. 1031–1044, 2010.
[11] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.
[12] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.
[13] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.
[14] M. J. Lai and J. Wang, "An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems,," *SIAM J. Optim.*, vol. 21, no. 1, pp. 82–101, 2011.
[15] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1 norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, 2006.
[16] L. Frank and J. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.
[17] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, Jul. 1995.
[18] J. Trzasko and A. Manduca, "Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization," *IEEE Trans. Med. Imag.*, vol. 28, no. 1, pp. 106–121, Jan. 2009.
[19] D. Malioutov and A. Aravkin, "Iterative log thresholding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7198–7202.
[20] J. H. Friedman, "Fast sparse regression and classification," *Int. J. Forecasting*, vol. 28, no. 3, pp. 722–738, 2012.
[21] C. Gao, N. Wang, Q. Yu, and Z. Zhang, "A feasible nonconvex relaxation approach to feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2014, pp. 356–361.
[22] X. Zhang, D. Wang, Z. Zhou, and Y. Ma, "Robust low-rank tensor recovery with rectification and alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 238–255, Jan. 2011.

[23] X. Zhang, Q. Liu, D. Wang, L. Zhao, N. Gu, and S. Maybank, "Self-taught semi-supervised dictionary learning with non-negative constraint," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 532–543, Jan. 2020.

[24] D. Wang, X. Zhang, M. Fan, and X. Ye, "Semi-supervised dictionary learning via structural sparse preserving," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2137–2144.

[25] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4130–4137.

[26] C. Lu, J. Tang, S. Yan, and Z. Lin, "Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 829–839, Feb. 2016.

[27] C. Lu, C. Zhu, C. Xu, S. Yan, and Z. Lin, "Generalized singular value thresholding," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1805–1811.

[28] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 37–45.

[29] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 217–224.

[30] A. Biswas and D. Jacobs, "Active image clustering with pairwise constraints from humans," *Int. J. Comput. Vis.*, vol. 108, no. 1/2, pp. 133–147, 2014.

[31] X. Luo, L. Zhang, F. Li, and B. Wang, "Graph embedding-based ensemble learning for image clustering," in *Proc. Int. Conf. Pattern Recognit.*, 2018, pp. 213–218.

[32] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5879–5887.

[33] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.

[34] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4838–4846.

[35] Q. Gao, S. Xu, C. Fang, C. Ding, X. Gao, and Y. Li, "R_1-2-DPCA and face recognition," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1212–1223, Apr. 2019.

[36] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu, "Sequential particle swarm optimization for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[37] R. Yang, Y. Zhu, X. Wang, C. Li, and J. Tang, "Learning target-oriented dual attention for robust RGB-T tracking," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3975–3979.

[38] H. Kieritz, W. Hubner, and M. Arens, "Joint detection and online multi-object tracking," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1459–1467.

[39] A. Rakotomamonjy, "Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms," *IEEE Trans. Signal Process.*, vol. 91, no. 7, pp. 1505–1526, Jul. 2011.

[40] W. Hu, W. Li, X. Zhang, and S. Maybank, "Single and multiple object tracking using a multi-feature joint sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 816–833, Apr. 2015.

[41] H. Wang, F. Nie, W. Cai, and H. Huang, "Semi-supervised robust dictionary learning via efficient $\ell_{2,0+}$ norms minimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1145–1152.

[42] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[43] Y. Xie, W. Zhang, Y. Qu, L. Dai, and D. Tao, "Hyper-laplacian regularized multilinear multiview self-representations for clustering and semisupervised learning," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 572–586, Feb. 2020.

[44] D. P. Bertsekas, *Nonlinear Programming*. Beijing, China: Tsinghua Univ. Press, 2018.

[45] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3047–3064.

[46] X. Li, J. Ren, S. Rambhatla, Y. Xu, and J. Haupt, "Robust PCA via dictionary based outlier pursuit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4699–4703.

[47] H. Zhang, Z. Lin, C. Zhang, and E. Y. Chang, "Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 3143–3149.

[48] C. Xiao, F. Nie, and H. Huang, "Exact top-k feature selection via $\ell_{2,0}$-norm constraint," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1240–1246.

[49] F. Nie, H. Huang, C. Xiao, and C. H. Q. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Neural Informat. Process. Syst.*, 2010, pp. 1–9.

[50] Y. Sun, Q. Liu, J. Tang, and D. Tao, "Learning discriminative dictionary for group sparse representation," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3816–3828, Sep. 2014.

[51] J. Yao, X. Liu, and C. Qi, "Foreground detection using low rank and structured sparsity," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2014, pp. 1–6.

[52] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.

[53] M. Ma, R. Hu, S. Chen, J. Xiao, and Z. Wang, "Robust background subtraction method via low-rank and structured sparse decomposition," *China Commun.*, vol. 15, no. 7, pp. 156–167, Jul. 2018.

[54] A. Zheng, Y. Zhao, C. Li, J. Tang, and B. Luo, "Moving object detection via robust low-rank and sparse separating with high-order structural constraint," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data*, 2018, pp. 1–6.

[55] J. Zhang and X. Jia, "Improved low rank plus structured sparsity and unstructured sparsity decomposition for moving object detection in satellite videos," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5421–5424.

[56] Y. Yang, Z. Yang, J. Li, and L. Fan, "Foreground-background separation via generalized nuclear norm and structured sparse norm based low-rank and sparse decomposition," *IEEE Access*, vol. 8, pp. 84 217–84 229, 2020.

[57] X. Shu, F. Porikli, and N. Ahuja, "Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3874–3881.

[58] J. Fan, L. Ding, Y. Chen, and M. Udell, "Factor group-sparse regularization for efficient low-rank matrix recovery," in *Proc. Adv. Neural Informat. Process. Syst.*, 2019, pp. 5104–5114.

[59] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 663–670.

[60] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2012.

[61] E. Ehsan and V. René, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[62] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[63] J. D. F. Habbema and J. Hermans, "Selection of variables in discriminant analysis by f-statistic and error rate," *Technometrics*, vol. 19, no. 4, pp. 487–493, 1977.

[64] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. Int. Workshop Mach. Learn.*, 1992, pp. 249–256.

[65] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[66] H. Qiao, P. Zhang, D. Wang, and B. Zhang, "An explicit nonlinear mapping for manifold learning," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 51–63, Feb. 2013.

[67] P. Zhou, C. Zhang, and Z. Lin, "Bilevel model-based discriminative dictionary learning for recognition," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1173–1187, Mar. 2017.

[68] L. Chen, X. Li, D. Sun, and K. C. Toh, "On the equivalence of inexact proximal ALM and ADMM for a class of convex composite programming," 2018, *arXiv:1803.10803*.

[69] Z. Zhou, *Machine Learning*. Beijing, China: Tsinghua Univ. Press, 2016.

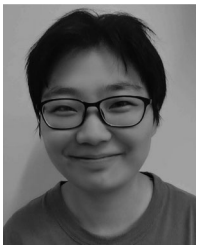
[70] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Int. Conf. Neural Informat. Process. Syst.*, 2010, pp. 1813–1821.

[71] M. Fan, X. Chang, X. Zhang, D. Wang, and L. Du, "Top-k supervise feature selection via ADMM for integer programming," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1646–1653.

[72] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[73] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.
[74] Z. Yu et al., "A new kind of nonparametric test for statistical comparison of multiple classifiers over multiple datasets," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4418–4431, Dec. 2017.
[75] E. M. Aharon Michal and B. Alfred, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[76] J. P. G. S. Julien Mairal, F. Bach, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[77] X. Lu, G. Tang, D. Wang, X. Zhang, and J. Zheng, "Structural dictionary learning based on non-convex surrogate of, norm for classification," in *Proc. IEEE Int. Conf. Big Data*, 2019, pp. 5056–5061.

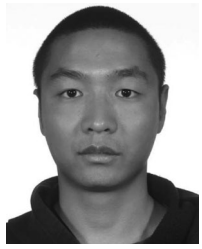
Xiaoqin Zhang (Senior Member, IEEE) received the BE degree in electronic information science and technology from Central South University, Changsha, China, in 2005, and the PhD degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently a professor with Wenzhou University, Wenzhou, China. He has authored or coauthored more than 80 papers in international and national journals and international conferences. His research interests include pattern recognition, computer vision, and machine learning.



Jingjing Zheng received the BA degree in art and design from the Wuchang Institute of Technology, Wuhan, China, in 2015, and the MS degree in applied mathematics from the College of Mathematics and Physics, Wenzhou University, Wenzhou, China, in 2020. She is currently working toward the PhD degree with the Department of Computer Science, Memorial University, St. John's, NL, Canada, and co-supervised by Prof. X. Zhang and Prof. X. Jiang. Her current research interests include pattern recognition and machine learning.

Di Wang received the BS degree from Shandong University, Jinan, China, and the PhD degree in applied mathematics from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2007 and 2012, respectively. He is currently an associate professor with Xi'an Jiaotong University, China. His current research interests include machine learning and Big Data analysis.

Guiying Tang received the BS degree in mathematics and applied mathematics from the Sichuan Normal University, China, in 2017 and the master's degree in basic mathematics from Wenzhou University, Wenzhou, China, in 2020. Her current research interests include image dehazing and machine learning.



Zhengyuan Zhou received the BE degree in electrical engineering and computer sciences and the BA degree in mathematics from the University of California at Berkeley, Berkeley, CA, USA, in 2012, and the PhD degree from the Information System Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA, USA, in summer 2019. He has substantial industry experience at Google, Microsoft Research, and Oracle. He is currently an Assistant Professor with the NYU Stern School of Business, New York, NY, USA. His research interests include learning, optimization, game theory, and stochastic systems.

Zhouchen Lin (Fellow, IEEE) received the PhD degree from Peking University, Beijing, China, in 2000. He is currently a professor with the Key Laboratory of Machine Perception, School of Intelligence Science and Technology, Peking University. His research interests include machine learning and numerical optimization. He is also a fellow of the International Association of Pattern Recognition (IAPR) and China Society of Image and Graphics (CSIG). He has been an area chair of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), Annual Conference on Neural Information Processing Systems (NIPS/NeurIPS), AAAI Conference on Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), International Conference on Learning Representations (ICLR), and International Conference on Machine Learning (ICML) many times. He was the program co-chair of International Conference on Pattern Recognition (ICPR) 2022 and the senior area chair of ICML 2022, NeurIPS 2022 and CVPR 2023. He is an associate editor of the International Journal of Computer Vision and Optimization Methods and Software.