Simultaneously Sparse and Low-Rank Matrix Reconstruction via Nonconvex and Nonseparable Regularization

Wei Chen, *Senior Member, IEEE*

***Abstract*—Many real-world problems involve the recovery of a matrix from linear measurements, where the matrix lies close to some low-dimensional structure. This paper considers the problem of reconstructing a matrix with a simultaneously sparse and low-rank model. As surrogate functions of the sparsity and the matrix rank that are non-convex and discontinuous, the ℓ_1 norm and the nuclear norm are often used instead to derive efficient algorithms to promote sparse and low-rank characteristics, respectively. However, the ℓ_1 norm and the nuclear norm are loose approximations, and furthermore, recent study reveals using convex regularizations for joint structures cannot do better, orderwise, than exploiting only one of the structures. Motivated by the construction of nonconvex and nonseparable regularization in sparse Bayesian learning, a new optimization problem is formulated in the latent space for recovering a simultaneously sparse and low-rank matrix. The newly proposed nonconvex cost function is proved to have the ability to recover a simultaneously sparse and low-rank matrix with a sufficient number of noiseless linear measurements. In addition, an algorithm is derived to solve the resulting non-convex optimization problem, and convergence analysis of the proposed algorithm is provided in the paper. The performance of the proposed approach is demonstrated by experiments using both synthetic data and real hyperspectral images for compressive sensing applications.**

I. Introduction

MATRIX reconstruction from a small number of linear measurements is at the center of many modern signal processing applications such as compressed sensing (CS) [1], [2] and matrix completion [3], [4]. If the matrix is arbitrary, there is indeed little we can do in this context. By making some assumptions about the matrix structure, then we can make some progress. For instance, we can assume the matrix has a low rank or sparse entries by applying some transformation. The core idea is that the number of degrees of freedom should be smaller than the number of entries in a "meaningful" discrete signal.

In various applications in signal processing and machine learning, the matrix of interest is known to have several structures at the same time [5]. One interesting assumption for matrices, which has been extracting much attention in the past years, is the simultaneously sparse and low-rank model. This simultaneously structured model has been exploited for instances in sub-wavelength optical imaging [6], hyperspectral image unmixing [7], [8] and feature coding [9].

Mathematically, the problem of recovering a simultaneously sparse and low-rank matrix from noiseless linear measurements can be described as

$$\begin{aligned} \min_{\mathbf{X}} \quad & \alpha\|\mathbf{X}\|_0 + (1-\alpha)\|\mathbf{X}\|_{\text{rank}} \\ \text{s.t.} \quad & \mathcal{A}[\mathbf{X}] = \mathbf{y}, \end{aligned} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the unknown matrix, $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^p$ is a linear mapping, $\mathbf{y} \in \mathbb{R}^p$ is a vector of measurements, and $0 \leq \alpha \leq 1$. Here $\|\mathbf{X}\|_0$ denotes the matrix ℓ_0 "norm" that counts the number of nonzero entries in $\mathbf{X}$, and $\|\mathbf{X}\|_{\text{rank}}$ denotes the matrix rank. The low-rank model and the sparse model capture the characteristics of the matrix from different perspectives. For a matrix with rank $r < \min\{n, m\}$, the number of degrees of freedom is $(n + m)r - r^2$, while the number of degrees of freedom of a k ($k < nm$) sparse matrix is k. On the other hand, let $\mathbf{X}$ be a rank r matrix whose entries are zero outside a $k_1 \times k_2$ submatrix. The number of degrees of freedom of the simultaneously sparse and low-rank matrix is only $(k_1 + k_2)r - r^2$. Therefore, it is expected that a smaller number of measurements is required to successfully reconstruct $\mathbf{X}$ by exploiting the joint structure.

Unfortunately, the non-convexity and discontinuous nature of the ℓ_0 "norm" and the rank function make the problem (1) challenging to solve. Efficient procedures developed in literature to deal with the sparse model and the low rank model mostly rely on convex relaxation, where the ℓ_0 "norm" and the rank function are replaced by the ℓ_1 norm and the nuclear norm (i.e., the sum of singular values) [4], [10], respectively. Then a convex optimization problem can be posed as

$$\begin{aligned} \min_{\mathbf{X}} \quad & \alpha\|\mathbf{X}\|_1 + (1-\alpha)\|\mathbf{X}\|_* \\ \text{s.t.} \quad & \mathcal{A}[\mathbf{X}] = \mathbf{y}, \end{aligned} \tag{2}$$

where $\|\mathbf{X}\|_*$ denotes the nuclear norm. If the true $\mathbf{X}$ is sparse and low-rank, improved reconstruction performance via solving (2) has been observed in various scenarios [7]–[9], in comparison to only considering the sparse model or the low rank model. However, a recent study [5] reveals a fundamental limitation in the formulation (2) that exploits convex regularization with random linear measurements. Specifically, it shows that using convex regularization for

This work is supported by the Natural Science Foundation of China (61671046); the Beijing Natural Science Foundation (4182050); the State Key Laboratory of Rail Traffic Control and Safety (RCS2018ZZ002) of Beijing Jiaotong University; Major projects of Beijing Municipal Science and Technology Commission under Grant No. Z181100003218010; NFSC Outstanding Youth under Grant 61725101; National key research and development program under Grant 2016YFE0200900.

Wei Chen is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China (e-mail: weich@bjtu.edu.cn).

joint structures can do no better, orderwise, than exploiting only one of the structures. For the case of simultaneously sparse and low-rank matrices, Oymak et al. prove that by using a nonconvex formulation, $\mathbf{X}$ can be recovered from $\mathcal{O}(r(k_1 + k_2)\log n)$ measurements[1], which is much smaller than the convex formulation combining the ℓ_1 norm and the nuclear norm [5].

Despite the attractive property of nonconvex regularization for the simultaneous sparse and low-rank model, there exist many sub-optimal local minima where optimization algorithms could be trapped. One popular nonconvex regularization for the sparse model is the ℓ_p pseudo-norm ($0 \leq p < 1$) [11], [12]. As an equivalent of the ℓ_p pseudo-norm for matrix, a smooth Schatten-p function is exploited to approximate the matrix rank in [13]. The ℓ_p pseudo-norm and many other penalty functions developed in literature have a common separable property, meaning a transformation of the function can be decomposed[2] as $f(\mathbf{z}) = \sum_i g(z_i)$. In contrast, nonconvex and nonseparable penalties have been proposed by Wipf et al. to induce sparsity and low-rank in [14] and [15], respectively. It is demonstrated that the objective functions with nonseparable penalties have fewer sub-optimal local minima under certain conditions of the mapping $\mathcal{A}$ [14], [15]. In [16], [17], more formulations of nonseparable regularization for the sparse model are provided and the limitation of separable penalties are further studied.

In this paper, a new optimization problem is formulated for recovering a simultaneously sparse and low-rank matrix. Instead of directly regularizing the matrix, by employing a Gaussian prior model with zero mean, an optimization problem in the latent space involving only hyperparameters, i.e., covariance in the statistic model, is proposed. The new cost function is nonconvex and has nonseparable penalties. Given the estimated covariance, the matrix can be recovered by applying the maximum a posteriori (MAP) point estimate, which has a closed-form expression. To give the rationale of the proposed cost function, it is proved that the global minima of the cost function produces the maximally sparse and low-rank solution in the noiseless case under certain conditions on the linear mapping $\mathcal{A}$. An iterative algorithm is developed to solve the newly proposed non-convex optimization problem and convergence analysis on the proposed algorithm is conducted. The superior performance of the proposed approach in comparison with state-of-the-art alternatives is demonstrated by extensive experiments on both synthetic data and real hyperspectral images in CS applications.

The rest of the paper is organized as follows: Section II describes related work on the simultaneously sparse and low-rank model and the sparse Bayesian learning (SBL) [18], [19] framework for sparse vector reconstruction. Section III provides the new nonconvex and nonseparable regularization for matrix reconstruction with the simultaneously sparse and low-rank model, and analysis on the global minimum of the cost function. In Section IV, an iterative algorithm is developed to solve the non-convex optimization problem and convergence analysis is conducted. Numerical results are presented in Section V, followed by conclusions in Section VI.

Throughout this paper, lower-case letters denote scalars, boldface upper-case letters denote matrices, and boldface lower-case letters denote column vectors. For a matrix $\mathbf{X}$, the superscripts $(\mathbf{X})^T$, $(\mathbf{X})^{-1}$, $(\mathbf{X})^\dagger$ and $|\mathbf{X}|$ denote the transpose, the inverse, the Moore-Penrose pseudoinverse and the determinant of $\mathbf{X}$, respectively. The trace of a matrix is denoted by Tr$[\cdot]$. rank$[\mathbf{X}]$ denotes the matrix rank. The operator vec$[\cdot]$ denotes vectorization for a matrix, and diag$[\cdot]$ denotes the diagonal vector of a matrix. $\mathbf{I}$ denotes an identity matrix, and $\mathbf{I}_p$ denotes a $p \times p$ identity matrix.

II. Background

This section introduces related work on the simultaneously sparse and low-rank model and the SBL [18], [19], which formulates nonconvex and nonseparable regularization for the sparse model from a Bayesian perspective.

A. Related Work on the Simultaneously Sparse and Low-rank Model

The sum of the sparse model and the low-rank model has already been considered in a different context. In robust principle component analysis (PCA) [20], a data matrix $\mathbf{X}$ is decomposed as $\mathbf{X} = \mathbf{S} + \mathbf{C}$, where $\mathbf{S}$ is sparse and $\mathbf{C}$ is low-rank. Then $\mathbf{X}$ is reconstructed by using ℓ_1 norm regularization over $\mathbf{S}$ and nuclear norm regularization over $\mathbf{C}$. Applications using this approach include image background modeling, dimensionality reduction [21], and covariance estimation [22].

In this paper, different to the robust PCA, the matrix $\mathbf{X}$ is simultaneously sparse and low-rank. One would like to come up with algorithms that exploit both types of structures to minimize the number of measurements required for recovery. In [23], Yang et al. show that an iterative thresholding algorithm achieves (near) optimal rates adaptively under mild conditions for matrix denoising applications where the mapping $\mathbf{A} = \mathbf{I}$. In [24]–[26], the ℓ_1 norm and the nuclear norm are employed to promote sparsity and low rank, respectively, which together with a data fidelity term, result in a convex optimization problem. The simultaneously sparse and low-rank model with convex approximated penalties has also been studied in a variety of scenarios, e.g., hyperspectral image unmixing [7], [8] and feature coding [9]. In [27], [28], a related but more strict model, i.e., the simultaneously row-sparse and low-rank model, is exploited for hyperspectral image compressed sensing, where the convex $\ell_{2,1}$ norm is used to approximate the row sparsity of a matrix. Another related work is sparse matrix factorization [29], which aims to infer a low-rank matrix that can be factorized as the product of two sparse matrices with few columns (left factor) and few rows (right factor).

The use of convex approximations enables one to apply well-developed convex optimization techniques to solve the problem, and conduct theoretical analysis on the error bound

[1] Here, $\mathbf{X}$ is considered as a square matrix, i.e., $m = n$. Note that in comparison to the number of degrees of freedom in $\mathbf{X}$ (i.e., $(k_1 + k_2)r - r^2$), the sampling complexity of using a nonconvex formulation (i.e., $\mathcal{O}(r(k_1 + k_2)\log n)$) is degraded only by a logarithmic factor.

[2] For instance, the ℓ_p pseudo-norm can be written as $f(\mathbf{z}) = \|\mathbf{z}\|_p^p = \sum_i |z_i|^p = \sum_i g(z_i)$.

and convergence. However, owing to the approximation, a convex approach will fail when the global minimum of the cost function is not equal to the true $\mathbf{X}$. There is nothing one can do to avoid this structural error for convex approaches from the perspective of algorithmic development. In [5], Oymak et al. investigate the number of linear measurements required to recover simultaneously structured models, and theoretically demonstrate that a combination of the ℓ_1 norm and the nuclear norm cannot perform significantly better than the best individual norm. Although it seems that using convex relaxation for simultaneously structured models is not promising, Oymak et al. further find the nonconvex recovery method[3] could benefit from both structures and is only slightly suboptimal in terms of sampling complexity in comparison to the number of degrees of freedom. In [30], a strictly convex objective function with nonconvex and separable penalties is proposed for estimating a sparse low-rank matrix from its noisy observation. The work in [30] focuses on the denoising case, and it is not clear how to extend to non-separable penalties.

B. Sparse Bayesian Learning

SBL [18], [19] is a popular approach for single sparse signal recovery from a Bayesian perspective and the resulting cost function benefits from nonconvex and nonseparable regularization. Consider a sparse vector that is observed as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{3}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{p\times n}$, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{e} \in \mathbb{R}^p$ denote the measurement vector, the measurement matrix, the unknown sparse vector to be estimated, and the noise vector, respectively. Then assume the likelihood $p(\mathbf{y}|\mathbf{x})$ to be Gaussian with noise variance λ, which is expressed as

$$p(\mathbf{y}|\mathbf{x};\lambda) = \mathcal{N}(\mathbf{A}\mathbf{x}, \lambda\mathbf{I}_p). \tag{4}$$

Furthermore, SBL considers a Gaussian prior model

$$p(\mathbf{x}|\mathbf{\Gamma}) = \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}), \tag{5}$$

where $\mathbf{\Gamma} \in \mathbb{R}^{n\times n}$ is a diagonal matrix.

In contrast to standard Bayesian methods, where the prior distribution is fixed before any data are observed, SBL estimates the prior distribution, i.e., $\mathbf{\Gamma}$, from the data $\mathbf{y}$ by applying the MAP estimation

$$\begin{aligned} \boldsymbol{\gamma} &= \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|\mathbf{y}) \\ &= \arg\max_{\boldsymbol{\gamma}} \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x};\boldsymbol{\gamma})d\mathbf{x} \\ &= \arg\min_{\boldsymbol{\gamma}} \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \log\left|\mathbf{\Sigma}\right|, \end{aligned} \tag{6}$$

where $\boldsymbol{\gamma} = \text{diag}[\mathbf{\Gamma}] \in \mathbb{R}^n$ and $\mathbf{\Sigma} = \lambda\mathbf{I} + \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T$. Given the likelihood (4) and prior (5), the posterior distribution $p(\mathbf{x}|\mathbf{y};\mathbf{\Sigma})$ is a Gaussian with mean

$$\mathbf{x} = \mathbf{\Gamma}\mathbf{A}^T(\lambda\mathbf{I} + \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T)^{-1}\mathbf{y}. \tag{7}$$

While SBL with the cost function in (6) may seem quite different to other methods that directly penalize on the signal $\mathbf{x}$, it can be reexpressed in the $\mathbf{x}$-space as solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}) \\ \text{s.t.} \quad & f(\mathbf{x}) = \min_{\boldsymbol{\gamma}\geq 0} \mathbf{x}^T\mathbf{\Gamma}\mathbf{x} + \log\left|\mathbf{\Sigma}\right|. \end{aligned} \tag{8}$$

It is proved in [14] that $\mathbf{x}^*$ is a local minimum of (8) if and only if $\boldsymbol{\gamma}^*$ is a local minimum of (6). Obviously, $f(\mathbf{x})$ given in (8) is nonseparable, meaning $f(\mathbf{x}) \neq \sum_i g(x_i)$. This nonseparable regularization term can be seen as an approximation that promotes sparsity. The benefit of using the nonseparable regularization given in (8) is that it produces fewer local minima than when using $\|\mathbf{x}\|_0$ directly, while separable regularization terms, e.g., the ℓ_p pseudo-norm, fail in this regard [14].

Similar to SBL that exploits empirical Bayes methods, Xin et al. formulate a new nonseparable regularization, namely BARM, for the low-rank model [15]. The following prior for a low rank matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ is used in [15]

$$p(\mathbf{X}|\mathbf{\Psi}) = \prod_i \mathcal{N}(\mathbf{x}_{:i}; \mathbf{0}, \mathbf{\Psi}), \tag{9}$$

where $\mathbf{\Psi} \in \mathbb{R}^{n\times n}$ is a positive semi-definite symmetric matrix. Then the matrix $\mathbf{X}$ can be estimated from linear measurements by using formal statistical inference procedures as SBL. However, it is not clear how to establish the simultaneously sparse and low-rank model from the same perspective. A similar problem occurs in our previous work [31], where we study a simultaneously row-sparse and element-sparse model, while unfortunately owing to the large difference between the row-sparse model and the low rank model of a matrix[4], the results in [31] cannot be directly applied to the simultaneously sparse and low-rank model.

III. Nonconvex and Nonseparable Regularization for the Simultaneously Sparse and Low-Rank Matrix Model

In this section, a new nonconvex and nonseparable regularization is proposed for matrix reconstruction with the simultaneously sparse and low-rank model. Then analysis on the global minimum of the cost function is conducted.

A. A New Cost Function

Define $\mathbf{X} \in \mathbb{R}^{n\times m}$ as an unknown matrix, and $\mathbf{A} \in \mathbb{R}^{p\times nm}$ as a matrix corresponding to the linear operator $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ such that the measurement vector is $\mathbf{y} = \mathbf{A}\mathbf{x}$. Let $\alpha > 0$, $\beta > 0$, $\mathbf{\Gamma} \in \mathbb{R}^{nm\times nm}$ be a diagonal matrix with the diagonal vector $\boldsymbol{\gamma} = \text{diag}[\mathbf{\Gamma}] \in \mathbb{R}^{nm}$, $\mathbf{\Psi} \in \mathbb{R}^{n\times n}$ be a positive semi-definite symmetric matrix[5], $\bar{\mathbf{\Psi}} = \mathbf{I}_m \otimes \mathbf{\Psi}$, and

$$\mathbf{\Phi}^{-1} = \mathbf{\Gamma}^{-1} + \bar{\mathbf{\Psi}}^{-1}. \tag{10}$$

[3] For nonconvex recovery, the theoretical analysis in [5] is based on properties of the global minimum of a nonconvex problem.

[4] The low-rank model and the row-sparse model capture the characteristics of the data from different perspectives.

[5] Technically, $\mathbf{\Psi}$ must be positive definite for invertibility. For the convenience throughout the paper with slight abuse of notation, the Moore-Penrose pseudoinverse is used as the "inverse" of a positive semi-definite matrix. With the eigendecomposition $\mathbf{\Psi} = \mathbf{Q}\mathbf{R}\mathbf{Q}^T$, the Moore-Penrose pseudoinverse is $\mathbf{\Psi}^\dagger = \mathbf{Q}\mathbf{R}^\dagger\mathbf{Q}^T$, where $\mathbf{R}^\dagger$ is computed by using the inverse of nonzero diagonal entries of $\mathbf{R}$ and setting others to zero.

By assuming a Gaussian likelihood as in (4), a Gaussian prior $p(\mathbf{x}|\boldsymbol{\Phi}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$ and an improper (unnormalisable) hyperprior probability density[6] $p(\boldsymbol{\gamma}, \boldsymbol{\Psi}) = \frac{|\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T|^{-\frac{\alpha}{2}}|\lambda\mathbf{I}+\mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T|^{-\frac{\beta}{2}}}{|\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T|^{-\frac{1}{2}}}$, and conducting MAP estimation $\max_{\boldsymbol{\gamma}\geq 0, \boldsymbol{\Psi}\succeq 0} \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}; \boldsymbol{\gamma}, \boldsymbol{\Psi})p(\boldsymbol{\gamma}, \boldsymbol{\Psi})d\mathbf{x}$, the following optimization problem is proposed for simultaneously sparse and low-rank matrix reconstruction

$$\min_{\substack{\boldsymbol{\gamma}\geq 0, \boldsymbol{\Psi}\succeq 0,\\ \bar{\boldsymbol{\Psi}}=\mathbf{I}_m\otimes\boldsymbol{\Psi},\\ \boldsymbol{\Phi}^{-1}=\boldsymbol{\Gamma}^{-1}+\bar{\boldsymbol{\Psi}}^{-1}}} \mathbf{y}^T(\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y} + \alpha\log|\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| + \beta\log|\lambda\mathbf{I}+\mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T|, \tag{11}$$

where the cost function is nonconvex and nonseparable. Given $\boldsymbol{\gamma}$ and $\boldsymbol{\Psi}$, i.e., the optimal solutions of (11), the reconstructed matrix with column-wise vectorization is

$$\mathbf{x} = \mathrm{vec}[\mathbf{X}] = \boldsymbol{\Phi}\mathbf{A}^T(\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y}. \tag{12}$$

The proposed optimization problem (11) in the latent space may seem quite different to the cost functions, e.g., (1) and (2), in the original $\mathbf{X}$ space. To shed light on the connection, a dual problem of (11) in the $\mathbf{X}$ space is developed in the sequel.

Now, by involving a new variable $\mathbf{x} \in \mathbb{R}^{nm}$, the first term in the cost function of (11) can be upper bounded by

$$\mathbf{y}^T(\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y} \leq \frac{1}{\lambda}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \mathbf{x}^T\boldsymbol{\Phi}^{-1}\mathbf{x}, \tag{13}$$

where the equality holds if (12) is satisfied. This upper bound can be proved by firstly minimizing the right-hand side of (13) with respect to the newly introduced variable $\mathbf{x}$, which leads to the optimal solution given in (12), and then inserting (12) back to the right-hand side of (13), which leads to the left-hand side of (13). By using this upper bound, a dual problem of (11) in the $\mathbf{X}$ space can be expressed as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}) \\ \text{s.t.} \quad & f(\mathbf{x}) = \min_{\substack{\boldsymbol{\gamma}\geq 0, \boldsymbol{\Psi}\succeq 0,\\ \bar{\boldsymbol{\Psi}}=\mathbf{I}_m\otimes\boldsymbol{\Psi},\\ \boldsymbol{\Phi}^{-1}=\boldsymbol{\Gamma}^{-1}+\bar{\boldsymbol{\Psi}}^{-1}}} \mathbf{x}^T\boldsymbol{\Phi}^{-1}\mathbf{x} + \alpha\log|\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| \\ & \qquad + \beta\log|\lambda\mathbf{I}+\mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T|. \end{aligned} \tag{14}$$

The relationship between the proposed optimization problem (11) in the latent space and its dual problem (14) in $\mathbf{X}$ space is given in the following theorem.

Theorem 1: Let $\hat{\mathbf{x}} = \hat{\boldsymbol{\Phi}}\mathbf{A}^T(\lambda\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y}$, $\hat{\boldsymbol{\Phi}}^{-1} = \hat{\boldsymbol{\Gamma}}^{-1} + \hat{\bar{\boldsymbol{\Psi}}}^{-1}$, and $\hat{\bar{\boldsymbol{\Psi}}} = \mathbf{I}_m \otimes \hat{\boldsymbol{\Psi}}$. Then $\hat{\mathbf{x}}$ is a global/local minimizer of (14) if and only if $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$ are global/local minimizers of (11).

Proof: We first prove that given global minimizers of (11), i.e., $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$, $\hat{\mathbf{x}} = \hat{\boldsymbol{\Phi}}\mathbf{A}^T(\lambda\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y}$ must be a global minimizer of (14). As $\hat{\mathbf{x}} = \hat{\boldsymbol{\Phi}}\mathbf{A}^T(\lambda\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y}$, the upper bound in (13) is tight, i.e.,

$$\mathbf{y}^T(\lambda\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y} = \frac{1}{\lambda}\|\mathbf{y}-\mathbf{A}\hat{\mathbf{x}}\|_2^2 + \hat{\mathbf{x}}^T\hat{\boldsymbol{\Phi}}^{-1}\hat{\mathbf{x}}. \tag{15}$$

Inserting (15) into (11) and removing unrelated terms, we have that $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$ are also global minimizers of

$$\min_{\substack{\boldsymbol{\gamma}\geq 0, \boldsymbol{\Psi}\succeq 0,\\ \bar{\boldsymbol{\Psi}}=\mathbf{I}_m\otimes\boldsymbol{\Psi},\\ \boldsymbol{\Phi}^{-1}=\boldsymbol{\Gamma}^{-1}+\bar{\boldsymbol{\Psi}}^{-1}}} \hat{\mathbf{x}}^T\boldsymbol{\Phi}^{-1}\hat{\mathbf{x}} + \alpha\log|\lambda\mathbf{I}+\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| + \beta\log|\lambda\mathbf{I}+\mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T|, \tag{16}$$

which suggests that $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$ are optimal solutions of the condition in (14). For fixed $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$, the unique optimal value of $\mathbf{x}$ in (14) is given by (12). Therefore, by construction, it follows $\hat{\mathbf{x}}$ is a global minimizer of (14).

Now we prove that if $\hat{\mathbf{x}} = \hat{\boldsymbol{\Phi}}\mathbf{A}^T(\lambda\mathbf{I}+\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y}$ is a global minimizer of (14), then $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$ are global minimizers of (11). As $\hat{\mathbf{x}}$ is a global minimizer of (14), we have $f(\hat{\mathbf{x}}) = \hat{\mathbf{x}}^T\hat{\boldsymbol{\Phi}}^{-1}\hat{\mathbf{x}} + \texttt{constant}$ in (14). Thus, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$ are optimal solutions of the constraint in (14) by construction. Inserting (15) into the optimization problem in the condition of (14) and removing unrelated terms, we have that $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Psi}}$ must be global minimizers of (11).

The relationship between global solutions to (11) and (14) extends to local optimal solutions, as given fixed $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$, the optimal $\mathbf{x}$ is unique in (12). ■

According to the Theorem 1, the two optimization problems are equivalent in terms of global/local minimizers. Obviously, $f(\mathbf{x})$ given in (14) is nonseparable. This dual space view of the original problem also helps in the derivation of the proposed algorithm for simultaneous sparse and low-rank model in Section IV.

B. Analysis on the Cost Function

The reconstructed matrix is linked via (12) with the proposed optimization problem (11) in the latent space. It would be good to provide the rationale why the proposed optimization problem (11) favors a solution that leads to the reconstructed $\mathbf{X}$ in (12) with a simultaneous sparse and low-rank structure.

Firstly, the log-determinant function is a concave non-decreasing function of the singular values of symmetric positive definite matrices, and thus the cost function (11) favors minimal rank of $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}$. As $\boldsymbol{\Gamma}$ is a diagonal matrix, a low-rank $\boldsymbol{\Gamma}$ leads to a sparse diagonal vector $\boldsymbol{\gamma}$. According to (10), the column space of $\boldsymbol{\Phi}$ is the intersection space between the column space of $\hat{\boldsymbol{\Psi}}$ and the column space of $\boldsymbol{\Gamma}$. Therefore, $\boldsymbol{\Phi}$ is low-rank and has only a few non-zero rows. As the reconstructed matrix $\mathbf{X}$ in (12) results from a left-multiplication with $\boldsymbol{\Phi}$, $\mathbf{X}$ must be simultaneous sparse and low-rank.

In the following formal result, it has been proved that the global minima of the cost function in (11) produces the maximally sparse and low-rank solution[7] in the noiseless case.

[6] For a sparse $\boldsymbol{\gamma}$ or a low rank $\boldsymbol{\Psi}$, we have $\lim_{\lambda\to 0} p(\boldsymbol{\gamma}, \boldsymbol{\Psi}) \to \infty$. Therefore, this improper hyperprior probability density promotes a simultaneously sparse and low-rank model.

[7] The level of simultaneously sparse and low-rank can be characterized as a multi-objective function of the sparsity and the rank associated with some weights.

Definition 1: The spark, spark$[\mathbf{A}]$, of a given matrix $\mathbf{A}$ is the smallest number of columns of $\mathbf{A}$ that are linearly dependent.

Theorem 2: Let $\mathbf{y} = \mathcal{A}[\mathbf{X}]$, where $\mathbf{X} \in \mathbb{R}^{n\times m}$ $(m \leq n)$, $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^{p}$ is a linear operator corresponding to a matrix $\mathbf{A} \in \mathbb{R}^{p\times nm}$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$. Define s and r as the sparsity level and rank of any feasible solution, respectively, that lead to the smallest $\alpha s + \beta m r$, where $\alpha > 0$ and $\beta > 0$. Then, if $s < p$, $r < \frac{p}{m}$ and spark$[\mathbf{A}] = p + 1$, in the limit $\lambda \rightarrow 0$, the global minima of (11) is achieved at $\hat{\gamma}$ and $\hat{\boldsymbol{\Psi}}$ such that $\hat{\mathbf{x}} = \hat{\boldsymbol{\Phi}}\mathbf{A}^T(\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y}$, where $\hat{\boldsymbol{\Phi}}^{-1} = \hat{\boldsymbol{\Gamma}}^{-1} + \hat{\bar{\boldsymbol{\Psi}}}^{-1}$ and $\hat{\bar{\boldsymbol{\Psi}}} = \mathbf{I}_m \otimes \hat{\boldsymbol{\Psi}}$.

Proof: In the limit $\lambda \rightarrow 0$, a minimizer of the cost function in (11) must satisfy $\mathbf{y} \in$ span$[(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{\frac{1}{2}}]$, otherwise the cost function would diverge to infinity as $\mathbf{y}(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y}$ tends to be infinity with a faster rate than the log-determinant terms approaching minus infinity[8]. The constraint $\mathbf{y} \in$ span$[(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{\frac{1}{2}}]$ is equivalent to requiring

$$\mathbf{y}^T(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y} \leq \rho,$$

where $\rho > 0$ denotes some finite bound.

While $\mathbf{y}(\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T)^{-1}\mathbf{y}$ is bounded, the minimum occurs when the log-determinant terms are approaching minus infinity. According to $\hat{\mathbf{x}} = \hat{\boldsymbol{\Phi}}\mathbf{A}^T(\mathbf{A}\hat{\boldsymbol{\Phi}}\mathbf{A}^T)^{-1}\mathbf{y}$ and $\hat{\boldsymbol{\Phi}}^{-1} = \hat{\boldsymbol{\Gamma}}^{-1} + \hat{\bar{\boldsymbol{\Psi}}}^{-1}$, we let $\|\hat{\gamma}\|_0 = s$ and rank$[\hat{\boldsymbol{\Psi}}] = r$. As $s < p$, $r < \frac{p}{m}$ and spark$[\mathbf{A}] = p + 1$, the sum of the log-determinant terms can be expressed by

$$\begin{aligned}
&\alpha \log|\lambda\mathbf{I} + \mathbf{A}\hat{\boldsymbol{\Gamma}}\mathbf{A}^T| + \beta \log|\lambda\mathbf{I} + \mathbf{A}\hat{\bar{\boldsymbol{\Psi}}}\mathbf{A}^T| \\
=&\alpha \sum_{i=1}^{p} \log|\lambda + \sigma_i[\mathbf{A}\hat{\boldsymbol{\Gamma}}\mathbf{A}^T]| + \beta \sum_{i=1}^{p} \log|\lambda + \sigma_i[\mathbf{A}\hat{\bar{\boldsymbol{\Psi}}}\mathbf{A}^T]| \\
=&\alpha \sum_{i=1}^{s} \log|\lambda + \sigma_i[\mathbf{A}\hat{\boldsymbol{\Gamma}}\mathbf{A}^T]| + \beta \sum_{i=1}^{mr} \log|\lambda + \sigma_i[\mathbf{A}\hat{\bar{\boldsymbol{\Psi}}}\mathbf{A}^T]| \\
&\quad + (\alpha(p-s) + \beta(p-mr)) \log|\lambda|,
\end{aligned} \tag{17}$$

where $\sigma_i[\cdot]$ denotes the ith singular value of a matrix. The first equality of (17) can be proved by using the singular value decomposition. Consequently, when $\lambda \rightarrow 0$, the sum of log-determinant terms scales as $(\alpha(p-s) + \beta(p-mr)) \log|\lambda|$, and hence the overall cost function is minimized when $\alpha s + \beta m r$ achieves its minimum. Now the proof is completed. ■

The condition spark$[\mathbf{A}] = p + 1$ in Theorem 2 can be satisfied almost surely by any random matrix with $p \leq nm$ [32]. This result explains why the proposed optimization problem is able to find exactly the true simultaneously sparse and low-rank matrix.

IV. Algorithm Derivation

In this section, an iterative algorithm is developed to solve the non-convex optimization problem in (11) for simultaneously sparse and low-rank matrix reconstruction, and convergence analysis on the proposed algorithm is provided.

A. Updating Rules

1) Update $\mathbf{x}$*:* As the optimization problem (11) is non-convex and hard to solve, the proposed algorithm uses majorization-minimization that repeatedly minimizes and updates surrogate function that majorizes the original cost function.

By exploiting the upper bound given in (13), the original optimization problem (11) can be solved by alternatively minimizing the following cost function using coordinate descent method:

$$\begin{aligned}
&\tilde{\mathcal{L}}(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\Psi}) \\
=&\frac{1}{\lambda}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mathbf{x}^T\boldsymbol{\Phi}^{-1}\mathbf{x} + \alpha \log|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| \\
&\quad + \beta \log|\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T|,
\end{aligned} \tag{18}$$

where the solution of $\min_{\mathbf{x}} \tilde{\mathcal{L}}(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\Psi})$ is given in (12). Therefore, by minimizing $\tilde{\mathcal{L}}(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\Psi})$, an updated estimation of the matrix $\mathbf{X}$ can be obtained according to (12).

It is worth mentioning that according to (10), the computation of $\boldsymbol{\Phi}$ involves a matrix inverse operation, which has a very high computational complexity $\mathcal{O}((nm)^3)$ and prohibits its application to recover large matrices. Fortunately, this computational complexity can be reduced to $\mathcal{O}(mn^3)$ by exploiting the block structure in $\boldsymbol{\Gamma}$ and $\bar{\boldsymbol{\Psi}}$. The ith diagonal sub-matrix of $\boldsymbol{\Phi}$ can be computed by

$$\begin{aligned}
\boldsymbol{\Phi}_i &= (\boldsymbol{\Gamma}_i^{-1} + \boldsymbol{\Psi}^{-1})^{-1} \\
&= \boldsymbol{\Gamma}_i - \boldsymbol{\Gamma}_i(\boldsymbol{\Psi} + \boldsymbol{\Gamma}_i)^{-1}\boldsymbol{\Gamma}_i,
\end{aligned} \tag{19}$$

where $\boldsymbol{\Gamma}_i \in \mathbb{R}^{n\times n}$ denotes the ith diagonal sub-matrix of $\boldsymbol{\Gamma}$, and the second equality is obtained by using the Woodbury identity on matrix inverse.

2) Update $\boldsymbol{\gamma}$ *and* $\boldsymbol{\Psi}$*:* As the log-determinant terms are concave nondecreasing functions, their concave conjugate functions can be defined as

$$h(\mathbf{z}) = \min_{\boldsymbol{\gamma}} \sum_{i=1}^{nm} \frac{z_i}{\gamma_i} - \log\left|\boldsymbol{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{A}^T\mathbf{A}\right|, \tag{20}$$

and

$$d(\mathbf{W}) = \min_{\substack{\boldsymbol{\Psi}\succeq 0 \\ \bar{\boldsymbol{\Psi}} = \mathbf{I}_m \otimes \boldsymbol{\Psi}}} \mathrm{Tr}[\mathbf{W}^T\boldsymbol{\Psi}^{-1}] - \log\left|\bar{\boldsymbol{\Psi}}^{-1} + \frac{1}{\lambda}\mathbf{A}^T\mathbf{A}\right|. \tag{21}$$

According to the duality relationship of concave conjugate functions, there are upper bounds

$$\log\left|\boldsymbol{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{A}^T\mathbf{A}\right| = \min_{\mathbf{z}} \sum_{i=1}^{nm} \frac{z_i}{\gamma_i} - h(\mathbf{z}), \tag{22}$$

and

$$\log\left|\bar{\boldsymbol{\Psi}}^{-1} + \frac{1}{\lambda}\mathbf{A}^T\mathbf{A}\right| = \min_{\mathbf{W}} \mathrm{Tr}[\mathbf{W}^T\boldsymbol{\Psi}^{-1}] - d(\mathbf{W}). \tag{23}$$

The bound in (22) is tight when the optimal value of z_i equals the slope at the current $\frac{1}{\gamma_i}$ of $\log\left|\boldsymbol{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{A}^T\mathbf{A}\right|$, i.e., $z_i = \nabla_{\frac{1}{\gamma_i}} \log\left|\boldsymbol{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{A}^T\mathbf{A}\right|$, which leads to

$$\mathbf{z} = \mathrm{diag}\left[\boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{A}^T\left(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right)^{-1}\mathbf{A}\boldsymbol{\Gamma}\right]. \tag{24}$$

[8] The derivative of the cost function in (11) tends to be minus infinity when the diagonal elements of $\boldsymbol{\gamma}$ and the singular values of $\boldsymbol{\Psi}$ are getting close to zeros.

Similarly, the bound in (23) is tight when

$$\mathbf{W} = \sum_{i=1}^{m} \boldsymbol{\Psi} - \boldsymbol{\Psi}\mathbf{A}_i^T \left(\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T\right)^{-1} \mathbf{A}_i\boldsymbol{\Psi}. \tag{25}$$

Inserting the upper bounds, (22) and (23), into the cost function (18) and omitting irrelevant terms, we arrive at the following approximation

$$\begin{aligned}\min_{\substack{\boldsymbol{\gamma},\boldsymbol{\Psi}\succeq 0\\ \bar{\boldsymbol{\Psi}}=\mathbf{I}_m\otimes\boldsymbol{\Psi}}} \quad & \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x} + \mathbf{x}^T\bar{\boldsymbol{\Psi}}^{-1}\mathbf{x} + \alpha\log|\boldsymbol{\Gamma}| + \beta m\log|\boldsymbol{\Psi}| \\ & + \alpha\sum_{i=1}^{nm}\frac{z_i}{\gamma_i} + \beta\mathrm{Tr}[\mathbf{W}^T\boldsymbol{\Psi}^{-1}],\end{aligned} \tag{26}$$

and its solutions are

$$\gamma_i = z_i + \frac{x_i^2}{\alpha} \tag{27}$$

and

$$\boldsymbol{\Psi} = \frac{1}{m}\left(\mathbf{W} + \frac{1}{\beta}\mathbf{X}\mathbf{X}^T\right). \tag{28}$$

Note that $\boldsymbol{\Psi}$ is positive semi-definite, if $\boldsymbol{\Psi}$ is initialized as a positive semi-definite symmetric matrix.

B. Convergence Analysis

By iteratively cycling through each of the above update rules, the proposed algorithm is derived for simultaneously sparse and low-rank matrix reconstruction, that are described in Algorithm 1. Although each iteration of the proposed algorithm is guaranteed to reduce or leave the cost function (11) unchanged, it is insufficient to guarantee formal convergence to a stationary point. Note that deriving a theoretical guarantee for an algorithm involving nonconvex and nonseparable regularization is usually very difficult, as it requires, for example, that the additional conditions of the Zangwills Global Convergence Theorem hold [33].

Here, the proposed algorithm can be guaranteed to converge to a stationary point with some modification to the original cost function (11). A small penalty term $\delta^{(t)}\left(\mathrm{Tr}[\boldsymbol{\Psi}^{-1}] + \mathrm{Tr}[\boldsymbol{\Gamma}^{-1}]\right)$ can be added to the cost function, where $\delta^{(t)} > 0$ is decreasing in each iteration of the algorithm. It is worth noting that the revised cost function is close to the original cost function when $\delta^t \to 0$. To incorporate the new penalty term, the proposed algorithm can be modified by replacing (27) and (28) with $\gamma_i = z_i + \frac{x_i^2}{\alpha} + \delta^{(t)}$ and $\boldsymbol{\Psi} = \frac{1}{m}\left(\mathbf{W} + \frac{1}{\beta}\mathbf{X}\mathbf{X}^T + \delta^{(t)}\right)$, respectively. Note that adding the small penalty term enables us to theoretically prove convergence, although this term does not need to be added in practice. Empirical evidences are provided to demonstrate the feasibility and applicability of the proposed algorithm in Section V with $\delta^{(t)} = 0$.

Theorem 3: Let $\{\delta^{(t)}\}$ be a decreasing positive sequence, and adjust the proposed algorithm to incorporate a new penalty term into (11), which leads to

$$\begin{aligned}\min_{\substack{\boldsymbol{\gamma}\geq 0,\boldsymbol{\Psi}\succeq 0,\\ \bar{\boldsymbol{\Psi}}=\mathbf{I}_m\otimes\boldsymbol{\Psi},\\ \boldsymbol{\Phi}^{-1}=\boldsymbol{\Gamma}^{-1}+\bar{\boldsymbol{\Psi}}^{-1}}} \quad & \mathbf{y}^T(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y} + \alpha\log|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| \\ & + \beta\log|\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T| + \delta^{(t)}\left(\mathrm{Tr}[\boldsymbol{\Psi}^{-1}] + \mathrm{Tr}[\boldsymbol{\Gamma}^{-1}]\right),\end{aligned} \tag{29}$$

Then the resulting sequence of iterations, i.e., $\{\boldsymbol{\gamma}^{(t)}, \boldsymbol{\Psi}^{(t)}\}$ is bounded, and every cluster point of the sequence is a stationary point of the optimization problem in (29).

Proof: To simplify the notation, we let $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\Psi}\}$, and define

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) =& \mathbf{y}^T(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Phi}\mathbf{A}^T)^{-1}\mathbf{y} + \alpha\log|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| \\ & + \beta\log|\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T| + \delta^{(t)}\left(\mathrm{Tr}[\boldsymbol{\Psi}^{-1}] + \mathrm{Tr}[\boldsymbol{\Gamma}^{-1}]\right)\end{aligned} \tag{30}$$

and

$$\begin{aligned}\tilde{\mathcal{J}}(\mathbf{x}, \boldsymbol{\theta}) =& \frac{1}{\lambda}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mathbf{x}^T\boldsymbol{\Phi}^{-1}\mathbf{x} + \alpha\log|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| \\ & + \beta\log|\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T| + \delta^{(t)}\left(\mathrm{Tr}[\boldsymbol{\Psi}^{-1}] + \mathrm{Tr}[\boldsymbol{\Gamma}^{-1}]\right).\end{aligned} \tag{31}$$

According to (13), we have $\mathcal{J}(\boldsymbol{\theta}) \leq \tilde{\mathcal{J}}(\mathbf{x}, \boldsymbol{\theta})$, where the equality holds if (12) is satisfied. Each iteration of the proposed algorithm can be guaranteed to reduce or leave the cost function $\mathcal{J}(\boldsymbol{\theta})$ unchanged, i.e., $\mathcal{J}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{J}(\boldsymbol{\theta}^{(t)})$ for all $t \geq 1$. To be specific, according to the update rule of $\mathbf{x}$, we have $\mathcal{J}(\boldsymbol{\theta}^{(t)}) = \tilde{\mathcal{J}}(\mathbf{x}^{(t+1)}, \boldsymbol{\theta}^{(t)})$. The update rule of $\boldsymbol{\theta}$ guarantees $\tilde{\mathcal{J}}(\mathbf{x}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}) \leq \tilde{\mathcal{J}}(\mathbf{x}^{(t+1)}, \boldsymbol{\theta}^{(t)})$. According to (13), we have $\mathcal{J}(\boldsymbol{\theta}^{(t+1)}) \leq \tilde{\mathcal{J}}(\mathbf{x}^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$. Therefore, the cost function is guaranteed to not increase, i.e., $\mathcal{J}(\boldsymbol{\theta}^{(t+1)}) \leq \mathcal{J}(\boldsymbol{\theta}^{(t)})$.

The derivative of the log-determinant term is upper bounded by

$$\begin{aligned}&\frac{\partial\alpha\log|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T|}{\partial\gamma_i} \\ =&\alpha\mathrm{Tr}[(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T)^{-1}\frac{\partial\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T}{\partial\gamma_i}] \\ =&\alpha\mathrm{Tr}[(\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T)^{-1}\mathbf{a}_i\mathbf{a}_i^T] \\ \leq&\alpha\mathrm{Tr}[(\lambda\mathbf{I})^{-1}\mathbf{a}_i\mathbf{a}_i^T],\end{aligned} \tag{32}$$

where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$. In addition, the derivative $\frac{\partial\mathrm{Tr}[\boldsymbol{\Gamma}^{-1}]}{\partial\gamma_i} \propto -\frac{1}{\gamma_i^2}$ tends to be minus infinity when $\frac{1}{\gamma_i} \to \infty$. Thus, $\alpha\log|\lambda\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T| + \delta^{(t)}\mathrm{Tr}[\boldsymbol{\Gamma}^{-1}]$ tends to be infinity with $\frac{1}{\gamma_i} \to \infty$. Similarly, it can be proved that $\beta\log|\lambda\mathbf{I} + \mathbf{A}\bar{\boldsymbol{\Psi}}\mathbf{A}^T| + \delta^{(t)}\mathrm{Tr}[\boldsymbol{\Psi}^{-1}]$ tends to be infinity when $\|\boldsymbol{\Psi}^{-1}\|_F \to \infty$. This observation leads to the conclusion that the sequence $\{\mathcal{J}(\boldsymbol{\theta}^{(t)})\}$ is lower bounded. Therefore, the sequence $\{\mathcal{J}(\boldsymbol{\theta}^{(t)})\}$ converges, which implies that the sequence of iterations $\{\boldsymbol{\theta}^{(t)}\}$ is bounded (as $\mathcal{J}(\boldsymbol{\theta}) \to \infty$ if and only if $\|\boldsymbol{\theta}\|_F \to \infty$).

Let $\tilde{\boldsymbol{\theta}}$ be a cluster point of $\{\boldsymbol{\theta}^{(t)}\}$, and suppose it is not a stationary point. According to the definition of cluster point, there exists a subsequence $\{\boldsymbol{\theta}^{(h)}\}$ of $\{\boldsymbol{\theta}^{(t)}\}$ converging to $\tilde{\boldsymbol{\theta}}$. By passing to a further subsequence if necessary, it can be assumed that $\{\boldsymbol{\theta}^{(h+1)}\}$ is convergent with a different limit $\breve{\boldsymbol{\theta}}$. By assumption, $\tilde{\boldsymbol{\theta}}$ is not a stationary point. Passing to limits, we see that $\tilde{\boldsymbol{\theta}}$ is not a minimizer of $\min_{\boldsymbol{\theta}}\tilde{\mathcal{J}}(\tilde{\mathbf{x}}, \boldsymbol{\theta})$, where $\tilde{\mathbf{x}} = \lim_{h\to\infty}\arg\min_{\mathbf{x}}\tilde{\mathcal{J}}(\mathbf{x}, \boldsymbol{\theta}^{(h)})$. Then we have $\tilde{\mathcal{J}}(\tilde{\mathbf{x}}, \breve{\boldsymbol{\theta}}) < \tilde{\mathcal{J}}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}})$, as $\boldsymbol{\theta}^{(h+1)}$ minimizes $\min_{\boldsymbol{\theta}}\tilde{\mathcal{J}}(\mathbf{x}^{(h)}, \boldsymbol{\theta})$. It follows that

$$\mathcal{J}(\breve{\boldsymbol{\theta}}) = \tilde{\mathcal{J}}(\breve{\mathbf{x}}, \breve{\boldsymbol{\theta}}) \leq \tilde{\mathcal{J}}(\tilde{\mathbf{x}}, \breve{\boldsymbol{\theta}}) < \tilde{\mathcal{J}}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\theta}}) = \mathcal{J}(\tilde{\boldsymbol{\theta}}). \tag{33}$$

Algorithm 1 The proposed algorithm for simultaneously sparse and low-rank matrix reconstruction

```
Step 1: Initialize $t = 0$, $\mathbf{\Gamma}^t = \mathbf{I}$, and $\mathbf{\Psi}^t = \mathbf{I}$;
Step 2: Compute $\mathbf{x}^t$ using (12);
Step 3: Compute $\mathbf{z}^t$ and $\mathbf{W}^t$ using (24) and (25), respectively;
Step 4: Compute $\mathbf{\Gamma}^{t+1}$ and $\mathbf{\Psi}^{t+1}$ using (27) and (28), respectively;
Step 5: Let $t = t+1$, and go to step 2 if some halting condition is not satisfied.
```

As the sequence $\{\mathcal{J}(\boldsymbol{\theta}^{(t)})\}$ converges, we have that

$$\lim_{t\to\infty} \mathcal{J}(\boldsymbol{\theta}^{(t)}) = \lim_{h\to\infty} \mathcal{J}(\boldsymbol{\theta}^{(h)}) = \mathcal{J}(\tilde{\boldsymbol{\theta}}) = \lim_{h\to\infty} \mathcal{J}(\boldsymbol{\theta}^{(h+1)}) = \mathcal{J}(\breve{\boldsymbol{\theta}}), \tag{34}$$

which contradict (33). Therefore, every cluster point of the sequence is a stationary point. ■

As the sequence $\{\mathcal{J}(\boldsymbol{\theta}^{(t)})\}$ is bounded, there is at least one cluster point $\boldsymbol{\theta}^*$ so that $\mathcal{J}(\boldsymbol{\theta}^*)$ is a stationary point. The algorithm might converge to a saddle point. However, this is rare [34] and any minimal perturbation leads to escape.

V. Numerical Experiments

In this section, the proposed algorithm is compared with the following approaches for simultaneously sparse and low-rank matrix reconstruction:

- Convex approach: using the CVX package [35] to solve the simultaneously sparse and low-rank matrix reconstruction problem with the convex ℓ_1 norm and nuclear norm regularization;
- Nonconvex approach [7]: using the state-of-the-art nonconvex approach with the weighted ℓ_1 norm and the weighted nuclear norm [7] for recovering a sparse and low-rank matrix;
- SBL [18], [19]: sparse vector reconstruction via nonconvex and nonseparable regularization;
- BARM [15]: low-rank matrix reconstruction via nonconvex and nonseparable regularization.

Reconstruction performance is evaluated by both synthetic data and real hyperspectral images for compressive sensing applications.

A. Experiments With Synthetic Data

In this subsection, a series of experiments with synthetic data are conducted in order to demonstrate the performance of the proposed simultaneously sparse and low-rank matrix reconstruction approach for different problem setups, i.e. different matrix sizes $n \times m$, different sparsity levels s, different ranks r and different number of measurements p. To generate the ground truth matrix $\mathbf{X}$, a nonsparse submatrix is produced by randomly choosing $\sqrt{s}$ rows and $\sqrt{s}$ columns of $\mathbf{X}$. All the elements not belonging to the submatrix are set to zeros. To enforce a low rank, the nonsparse submatrix $\mathbf{X}_s$ is generated as $\mathbf{X}_s = \mathbf{X}_{s,1}\mathbf{X}_{s,2} \in \mathbb{R}^{\sqrt{s}\times\sqrt{s}}$, where $\mathbf{X}_{s,1} \in \mathbb{R}^{\sqrt{s}\times r}$ and $\mathbf{X}_{s,2} \in \mathbb{R}^{r\times\sqrt{s}}$ are random matrices generated by independent and identically distributed Gaussian $\mathcal{N}(0,1)$. The entries of the sensing matrix $\mathbf{A} \in \mathbb{R}^{p\times mn}$ are generated independently from $\mathcal{N}(0,1)$. The recovery performance is evaluated via relative recovery error defined by $\frac{\|\hat{\mathbf{X}}-\mathbf{X}\|_F}{\|\mathbf{X}\|_F}$, and averaged over 100 trials. For the noiseless case, if the relative recover error is smaller than 10^{-3}, $\hat{\mathbf{X}}$ is regarded as a successful recovery of $\mathbf{X}$.

If not pointed out specifically in the experiments, the baseline settings in the simulation are given as: the number of measurements $p = 200$, the matrix dimension $m = n = 50$, the sparsity level $s = 100$, and the matrix rank $r = 4$. $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$ are simply fixed for the proposed algorithm to balance the sparsity and the low-rank model. For the convex approach and the nonconvex approach [7], the regularization parameters are fine-tuned with ten different values, i.e., $\{10^{-10}, 10^{-9}, \ldots, 10^{-1}\}$.

1) Noiseless Case: The first experiment studies how the proposed algorithm performs in the noiseless case to exactly recover the true matrix. With the default settings, Fig. 1 illustrates the the original simultaneously sparse and low-rank matrix in one trial and matrices reconstructed by various algorithms, where non-zero entries are in black. It is observed that the proposed algorithm is the only one that is able to successfully reconstruct the original simultaneously sparse and low-rank matrix.

More comparison results are shown in Fig. 2, where the number of measurements, the matrix dimension, the sparsity level and the matrix rank are varied in different sub-figures. For all the compared algorithms, the proposed algorithm with nonconvex and nonseparable regularization consistently achieves the best performance in all settings. For instance, according to Fig. 2 (d), the proposed algorithm could successfully recover matrices with rank $r = 5$, while all the other algorithms may fail when the matrix rank is greater than 1. The nonconvex approach [7] achieves better performance than the convex approach. Note that although the cost function in [7] is nonconvex, its sparse-enforcing penalty and low-rank-enforcing penalty are both separable, while the proposed approach in this paper employs nonseparable regularization. Limitations of separable penalties have been observed in [16], [17], and [14], [15] demonstrate that objective functions with nonseparable penalties lead to fewer sub-optimal local minima under certain conditions of the linear mapping. The SBL and the BARM only exploit either the sparse structure or the low-rank structure, and thus have significant performance degradation in comparison to the proposed algorithm.

We now evaluate the computing time of the proposed method. Our simulations are performed in a MATLAB R2014a environment on a system with a dual-core 3.4 GHz CPU and 16 GB RAM, running under the Microsoft Windows 7 operating system. As shown in Table I, the proposed algorithm takes more computing time than the nonconvex approach [7]. However, it could be argued that the proposed algorithm will be preferred in applications where a high-accuracy solution is desired.

Fig. 3 shows the convergence performance of the proposed algorithm. For the default settings, the algorithm converges with 250 iterations, while less iterations are required for

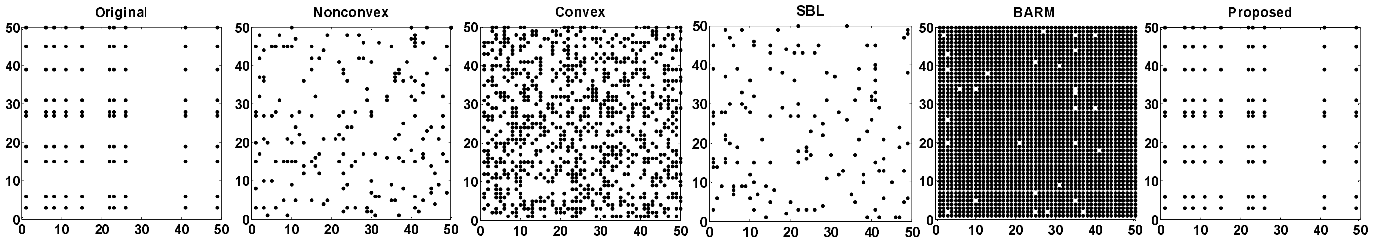
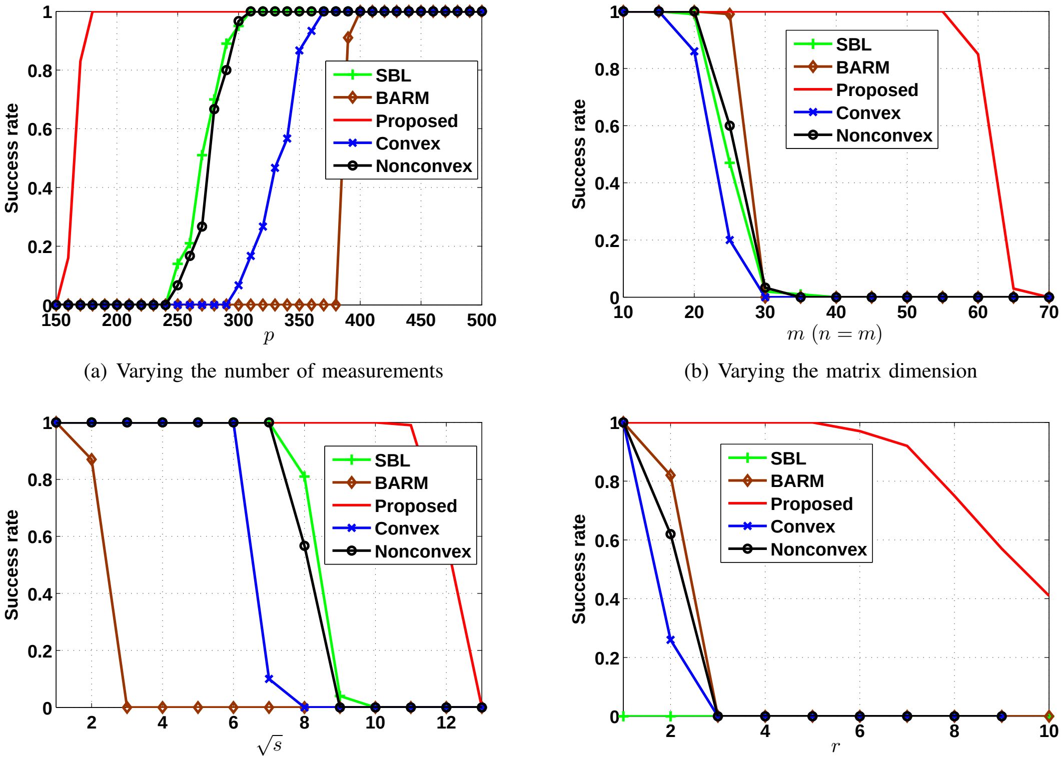


Fig. 1. Comparison of various algorithms for matrix reconstruction with the default settings.



(a) Varying the number of measurements

(b) Varying the matrix dimension

(c) Varying the sparsity level

(d) Varying the matrix rank

Fig. 2. Comparison of reconstruction success rate in the noiseless case.

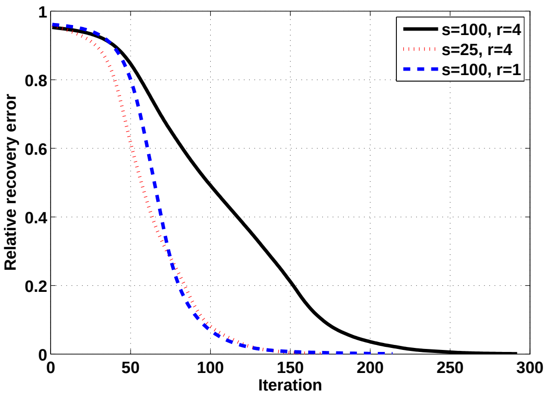


Fig. 3. Convergence rates of the proposed algorithm for a single instance.

relatively easy cases, i.e., ($s = 25$, $r = 4$) and ($s = 100$, $r = 1$). The computational complexity in each iteration of the proposed algorithm is $\mathcal{O}(p^3)$, which is the same as the SBL and the BARM.

2) Noisy Case: This experiment investigates how the proposed algorithm performs if the data is corrupted by noise. The ground truth matrix is randomly generated as the previous experiments in the noiseless case. Then an additive noise matrix is produced, where elements are generated following a zero-mean Gaussian distribution with variance adjusted to have a desired value of the signal to noise ratio (SNR). Results are shown in Fig. 4, where the proposed algorithm exhibits superior reconstruction accuracy in comparison to all the competitors when $p = 200$. One would observe that the performance of the state-of-the-art nonconvex approach [7] with $p = 400$ measurements and fine-tuned regularization parameters is close to the proposed algorithm. However, the proposed algorithm uses fixed regularization parameters, i.e., $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$, while its performance may be improved by fine-tuning α and β.

TABLE I
COMPARISON OF COMPUTING TIME (IN SECONDS)

Number of measurements p	Matrix dimension	Performance	SBL	BARM	Convex	Nonconvex	Proposed
200	50×50	Relative recovery error	6.9×10^{-1}	9.8×10^{-1}	7.7×10^{-1}	8.2×10^{-1}	9.9×10^{-4}
		Computing time	23.1	57.7	28.4	18.8	40.2
400	50×50	Relative recovery error	2.5×10^{-4}	9.9×10^{-4}	3.2×10^{-1}	1.4×10^{-3}	9.7×10^{-4}
		Computing time	13.0	38.9	71.3	20.8	25.6
400	100×100	Relative recovery error	3.6×10^{-4}	1.0	5.0×10^{-1}	9.4×10^{-3}	10×10^{-4}
		Computing time	260.4	923.9	503.8	315.7	644.6
600	50×50	Relative recovery error	3.8×10^{-5}	8.8×10^{-4}	1.1×10^{-5}	3.8×10^{-4}	8.8×10^{-4}
		Computing time	25.0	24.4	129.4	17.1	25.4
600	100×100	Relative recovery error	7.8×10^{-5}	0.4688	7.4×10^{-2}	5.3×10^{-3}	9.8×10^{-4}
		Computing time	281.6	1375	882.7	319.1	627.0

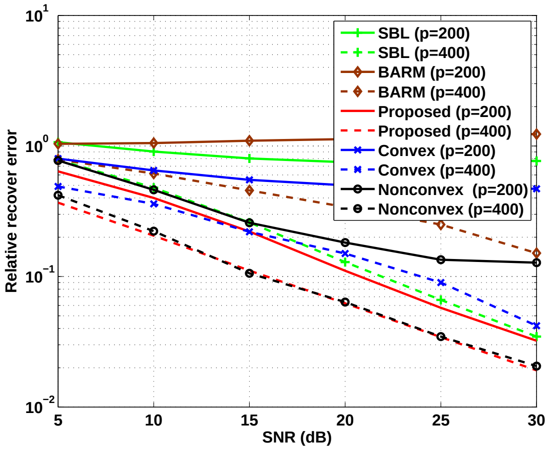
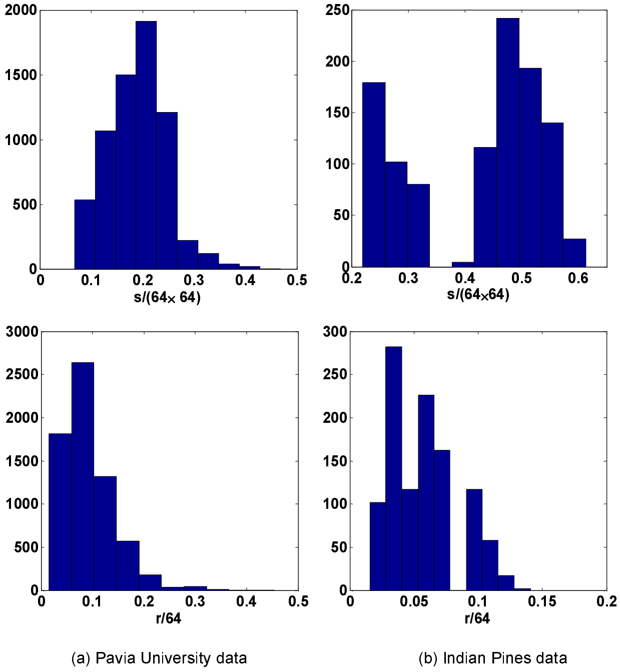


Fig. 4. Comparison of reconstruction accuracy in the noisy case.



(a) Pavia University data

(b) Indian Pines data

Fig. 5. Distributions of the sparsity level and the matrix rank for blocks of different hyperspectral data.

B. *Experiments With Real Hyperspectral Images*

Now the proposed algorithm is evaluated by considering compressive hyperspectral image reconstruction. Two hyperspectral datasets, i.e., the Pavia University scene and the Indian Pines scene, are used in the experiments. The Pavia University data was acquired by a flight campaign over Pavia in Italy and consists of 640×340 pixels and 103 spectral reflectance bands, while the Indian Pines data was gathered over the Indian Pines test site in north-western Indiana in US and consists of 145×145 pixels and 200 spectral bands. Given the size of the datasets, it will take a long time to reconstruct the full dataset by solving the inverse CS problem, which limits the effectiveness in practical CS hyperspectral imaging systems. One potential method to deal with this issue is to partition the scene into multiple sub-blocks and perform reconstruction algorithm in parallel for each block simultaneously. This strategy is employed in the experiments, where each hyperspectral data is partitioned into equally sized blocks with 8×8 pixels and 64 spectral bands. By vectorizing the 3D data cube in the spatial domain, a set of 64×64 matrices is obtained. For these matrices, the distribution of the matrix rank r and the distribution of the sparsity level[9] s under 1D 4 level discrete wavelet transform (DWT) is shown in Fig. 5. It is observed that the Pavia University data is more sparse than the Indian Pines data, while the Indian Pines data has a lower rank than the Pavia University data in the average case.

It is also assumed that a whiskbroom scanner is used to collect and compress each block pixel by pixel [36], then each hyperspectral block is recovered independently. Without loss of generality, the entries of the sensing matrix are generated independently from $\mathcal{N}(0,1)$. For the convex approach and the nonconvex approach [7], the regularization parameters are fine-tuned with ten different values, i.e., $\{10^{-10}, 10^{-9}, \ldots, 10^{-1}\}$, on a randomly selected subset of the data.

The reconstruction performance of various approaches is reported in Table I. For both hyperspectral datasets and different CS undersampling ratios, the proposed approach consistently outperforms all the competitors. The SBL that only exploits the sparse model has a worse reconstruction accuracy than the BARM that considers the low-rank model for hyperspectral datasets. Note that $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$ is simply set for the proposed algorithm to balance the sparsity and the low-rank model, and it is able to benefit from the joint model and

[9] The matrix rank is defined as the number of singular values which is greater than 1% of the largest singular value; similarly, the sparsity level of a matrix is defined as the number of entries whose absolute value is greater than 1% of the largest absolute value of all entries.

TABLE II
COMPARISON OF RECONSTRUCTION ACCURACY FOR HYPERSPECTRAL IMAGES

(a) Relative Recovery Error for Pavia University Data

Undersampling Ratio	SBL	BARM	Convex	Nonconvex	Proposed
10%	0.9894	0.9334	0.4003	0.4067	0.3553
20%	0.4440	0.0939	0.0817	0.0936	0.0508
30%	0.1027	0.0217	0.0332	0.0770	0.0191
40%	0.0659	0.0168	0.0224	0.0697	0.0148
50%	0.0411	0.0144	0.0168	0.0686	0.0130

(b) Relative Recovery Error for Indian Pines Data

Undersampling Ratio	SBL	BARM	Convex	Nonconvex	Proposed
10%	0.9336	0.9320	0.2468	0.1717	0.1306
20%	0.6781	0.0598	0.1325	0.0557	0.0368
30%	0.2526	0.0225	0.0714	0.0344	0.0207
40%	0.1735	0.0202	0.0213	0.0213	0.0182
50%	0.1288	0.0158	0.0163	0.0164	0.0147

achieves the best performance of all. It is envisaged that by adjusting α and β, the proposed algorithm may have improved reconstruction performance, while optimizing the parameter values require additional knowledge of the data, which is out of the scope this paper.

VI. CONCLUSION

In this paper, a novel optimization problem is proposed for simultaneously sparse and low-rank matrix reconstruction via nonconvex and nonseparable regularization. Distinct to traditional approaches that directly regularize the original signal, the new optimization problem is formulated in the latent space akin to the SBL that considers the sparse model. Theoretical analyses are provided to demonstrate the capability of the proposed nonconvex cost function to recover a simultaneously sparse and low-rank matrix. An algorithm is derived to solve the proposed optimization problem with convergence analysis. The superiority of the proposed approach has been demonstrated by experiments with synthetic data and also experiments involving hyperspectral images.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[4] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, June 2010.

[5] S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2886–2908, May 2015.

[6] Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev, "Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing," *Optics express*, vol. 19, no. 16, pp. 14 807–14 822, 2011.

[7] P. V. Giampouras, K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, "Simultaneously sparse and low-rank abundance matrix estimation for hyperspectral image unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4775–4789, Aug 2016.

[8] C. G. Tsinos, A. A. Rontogiannis, and K. Berberidis, "Distributed blind hyperspectral unmixing via joint sparsity and low-rank constrained non-negative matrix factorization," *IEEE Transactions on Computational Imaging*, vol. 3, no. 2, pp. 160–174, June 2017.

[9] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2429–2443, June 2016.

[10] M. Fazel, "Matrix rank minimization with applications," 2002.

[11] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on l_p-norm minimization," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 09, p. L09003, 2009.

[12] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[13] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *Journal of Machine Learning Research*, vol. 13, no. Nov, pp. 3441–3473, 2012.

[14] D. Wipf, B. Rao, and S. Nagarajan, "Latent variable bayesian models for promoting sparsity," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6236–6255, Sept 2011.

[15] B. Xin, Y. Wang, W. Gao, and D. Wipf, "Exploring algorithmic limits of matrix rank minimization under affine constraints," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 4960–4974, Oct 2016.

[16] I. W. Selesnick and I. Bayram, "Enhanced sparsity by non-separable regularization," *IEEE Transactions on Signal Processing*, vol. 64, no. 9, pp. 2298–2313, May 2016.

[17] I. Selesnick and M. Farshchian, "Sparse signal approximation via nonseparable regularization," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2561–2575, May 2017.

[18] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[19] D. Wipf and B. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, Aug 2004.

[20] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[21] G. Mateos and G. B. Giannakis, "Robust pca as bilinear decomposition with outlier-sparsity regularization," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5176–5190, Oct 2012.

[22] K. Greenewald and A. O. Hero, "Robust kronecker product pca for spatio-temporal covariance estimation," *IEEE Transactions on Signal Processing*, vol. 63, no. 23, pp. 6368–6378, Dec 2015.

[23] D. Yang, Z. Ma, and A. Buja, "Rate optimal denoising of simultaneously sparse and low rank matrices," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3163–3189, 2016.

[24] E. Richard, P.-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 51–58.

[25] S. Tariyal, H. K. Aggarwal, and A. Majumdar, "Hyperspectral impulse denoising with sparse and low-rank penalties," in *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, June 2015, pp. 1–4.

[26] T. C. Gélvez, H. F. Rueda, and H. Arguello, "Simultaneously sparse and low-rank hyperspectral image recovery from coded aperture compressive measurements via convex optimization," in *Algorithms and Technologies*

for Multispectral, Hyperspectral, and Ultraspectral Imagery XXII, vol. 9840. International Society for Optics and Photonics, 2016, p. 98401J.

[27] A. Gogna, A. Shukla, H. K. Agarwal, and A. Majumdar, "Split bregman algorithms for sparse / joint-sparse and low-rank signal recovery: Application in compressive hyperspectral imaging," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 1302–1306.

[28] M. Golbabaee and P. Vandergheynst, "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2741–2744.

[29] E. Richard, G. R. Obozinski, and J.-P. Vert, "Tight convex relaxations for sparse matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 3284–3292.

[30] A. Parekh and I. W. Selesnick, "Improved sparse low-rank matrix estimation," *Signal Processing*, vol. 139, pp. 62–69, 2017.

[31] W. Chen, D. Wipf, Y. Wang, Y. Liu, and I. J. Wassell, "Simultaneous bayesian sparse approximation with structured sparse models," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6145–6159, Dec 2016.

[32] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, April 2009.

[33] W. I. Zangwill, *Nonlinear programming: a unified approach*. Prentice-Hall Englewood Cliffs, NJ, 1969, vol. 196, no. 9.

[34] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 187–200, Jan 1999.

[35] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2013.

[36] R. M. Willett, M. F. Duarte, M. A. Davenport, and R. G. Baraniuk, "S-parsity and structure in hyperspectral imaging : Sensing, reconstruction, and target detection," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 116–126, Jan 2014.

Wei Chen (M'13-SM'18) received the B.Eng. degree and M.Eng. degree in Communications Engineering from Beijing University of Posts and T-elecommunications, China, in 2006 and 2009, respectively, and the Ph.D. degree in Computer Science from the University of Cambridge, UK, in 2013. Later, he was a Research Associate with the Computer Laboratory, University of Cambridge from 2013 to 2016. He is currently a Professor with Beijing Jiaotong University, Beijing, China. Dr. Chen is the recipient of the 2013 IET Wireless Sensor Systems Premium Award and the 2017 International Conference on Computer Vision (ICCV) Young Researcher Award. His current research interests include sparse representation, Bayesian inference, wireless communication systems and image processing.