NONCONVEX MIXED MATRIX MINIMIZATION

MEIJUAN SHANG
Department of Mathematics, School of Science
Shijiazhuang University, Shijiazhuang 050035, China

YANAN LIU AND LINGCHEN KONG
Department of Applied Mathematics, School of Science
Beijing Jiaotong University, Beijing 100044, China

XIANChAO XIu∗ AND YING YANG
State Key Lab for Turbulence and Complex Systems
Department of Mechanics and Engineering Science, College of Engineering
Peking University, Beijing 100871, China

(Communicated by Tianbao Yang)

ABSTRACT. Given the ultrahigh dimensionality and the complex structure, which contains matrices and vectors, the mixed matrix minimization becomes crucial for the analysis of those data. Recently, the nonconvex functions such as the smoothly clipped absolute deviation, the minimax concave penalty, the capped \( \ell_1 \)-norm penalty and the \( \ell_p \) quasi-norm with \( 0 < p < 1 \) have been shown remarkable advantages in variable selection due to the fact that they can overcome the over-penalization. In this paper, we propose and study a novel nonconvex mixed matrix minimization, which combines the low-rank and sparse regularizations and nonconvex functions perfectly. The augmented Lagrangian method (ALM) is proposed to solve the nonconvex mixed matrix minimization problem. The resulting subproblems either have closed-form solutions or can be solved by fast solvers, which makes the ALM particularly efficient. In theory, we prove that the sequence generated by the ALM converges to a stationary point when the penalty parameter is above a computable threshold. Extensive numerical experiments illustrate that our proposed nonconvex mixed matrix minimization model outperforms the existing ones.

1. Introduction. Modern technologies are producing a wealth of data with complex structures. For instance, in two-dimensional digital imaging, flow cytometry, and electroencephalography, matrix type covariates frequently arise when measurements are obtained for each combination of two underlying variables. To address scientific questions arising from those data, Zhou and Li [31] proposed the mixed matrix (MM) regression

\[
y = \langle X, B \rangle + \langle z, \gamma \rangle + \varepsilon,
\]  

(1)

2010 Mathematics Subject Classification. Primary: 90C26, 90C46; Secondary: 90C90, 65K10.

Key words and phrases. Complex structure, mixed matrix minimization, nonconvex optimization, augmented Lagrangian method.

This work was supported in part by the National Natural Science Foundation of China (61633001, 11671029, 11601348).

∗ Corresponding author: Xianchao Xiu.
where $y \in \mathbb{R}$ is the response, $X \in \mathbb{R}^{m \times q}$ is the design matrix and $z \in \mathbb{R}^p$ is the covariate vector, $B \in \mathbb{R}^{m \times q}, \gamma \in \mathbb{R}^p$ are the unknown regression coefficients, $\varepsilon \in \mathbb{R}$ is the noise which follows a normal distribution with mean 0 and standard deviation $\sigma$. Now the MM is the most popular approach for the data with complex structures. Next, we will address the recent developments in this direction as our approach is aligned with these developments.

1.1. Overview of the problem. In fact, if $B = 0$, model (1) will be reduced into the traditional linear regression

$$y = \langle z, \gamma \rangle + \varepsilon.$$  

As we all know that many researchers have made lots of contributions related to the theory, algorithms as well as applications, including the Lasso [21], the fused Lasso [22], the elastic net [30] and so on. If $\gamma = 0$, model (1) will be reduced into the traditional matrix regression

$$y = \langle X, B \rangle + \varepsilon.$$  

To the best of our knowledge, it can be traced back to earlier literatures by [15], and then developed to the nuclear norm [27], multivariate group Lasso [17], just to name a few.

To take the structure of $B$ into consideration, Zhou and Li [31] studied the following low-rank regularized mixed matrix (LMM) minimization

$$\min_{B, \gamma} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle - \langle z_i, \gamma \rangle)^2 + \mu \|B\|_*, \quad (2)$$  

in which $\|B\|_*$ is the nuclear norm (defined as the sum of all singular values) of $B$, and $\mu > 0$ is a tuning parameter. In statistics, they discussed the degrees of freedom. In algorithm, they established the Nesterov optimal gradient method for spectral regularized matrix regression model (2). However, they ignored the sparsity of $\gamma$, which is also natural in practice. Thus, Shang et al. [20] considered the low-rank and sparse regularized mixed matrix (LSMM) minimization

$$\min_{B, \gamma} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle - \langle z_i, \gamma \rangle)^2 + \mu \|B\|_* + \nu \|\gamma\|_1, \quad (3)$$  

in which $\|\gamma\|_1$ is the $\ell_1$ norm in componentwise (defined as the sum of absolute values of all entries). They discussed its degree of freedom, but didn’t give optimization algorithm.

Recently, the nonconvex penalty functions, such as the hard thresholding penalty [8], the $\ell_p$ penalty with $0 < p < 1$ [4], the smoothly clipped absolute deviation (SCAD) [7], the minimax concave penalty (MCP) [28], the capped $\ell_1$ penalty [29], have attracted much attention. It is well known that the $\ell_0$ norm is the essential measure of sparsity, and the $\ell_1$ norm is the best convex relaxation form. However, the $\ell_1$ norm usually induces over-penalization. All these efforts convince us that nonconvex penalty has nice nearly unbiased property, and could overcome the drawbacks of the $\ell_1$ norm [9]. Hence, the $\ell_1$ norm can be replaced by some functions of the nonconvex penalty if a more sparse solution is expected to be obtained with better reconstruction performance. Numerical experiments also illustrate that the nonconvex regularized model always offers many possibilities, see, e.g., signal processing [24, 25], image reconstruction [3, 16], foreground extraction [26, 32].
By integrating the mixed matrix regression and the nonconvex penalty functions, in this paper, we consider the nonconvex mixed matrix (NonMM) minimization

$$\min_{B, \gamma} \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle - \langle z_i, \gamma \rangle)^2 + \mu \|B\|_* + \phi(\gamma),$$

where $\phi$ is a nonconvex function of our interest, and the parameter $\nu$ is contained.

1.2. Contributions. Compared to previous work, the main innovations of our paper are summarized below:

1) We propose a novel nonconvex mixed matrix minimization model, which combines low-rank and sparse regularizations and nonconvex functions perfectly. To the best of our knowledge, we are the first to model the mixed matrix minimization in such a nonconvex framework.

2) In algorithm, we develop an efficient augmented Lagrange method (ALM) to optimize our proposed model, and the resulting subproblems either have closed-form solutions or can be solved by fast solvers. We also show that the sequence generated by the proposed ALM converges to a stationary point when the penalty parameter is chosen above a computable threshold.

3) Moreover, we conduct numerical experiments, including simulation data and real-world data, to demonstrate that our proposed nonconvex mixed matrix minimization model can achieve better performance than the existing state-of-the-art ones.

The remaining of this paper is organized as follows. In Section 2, some notations and preliminaries will be reviewed. The optimization algorithm and convergence result will be discussed in Section 3. In Section 4, extensive experiments will be conducted to substantiate the superiority of the proposed model over the other existing ones. This paper will be concluded in Section 5.
where $\hat{\partial} f(u)$ denotes the Fréchet subdifferential of $f$ at $u \in \text{dom} f$, which is the set of all $d \in \mathbb{R}^m$ satisfying
$$
\liminf_{y \neq u, y \to u} \frac{f(y) - f(u) - \langle d, y - u \rangle}{\|y - u\|} \geq 0.
$$

From the above definition, we can easily observe that
$$
\left\{ d \in \mathbb{R}^m : \exists x^k \to x, d^k \to d, d^k \in \partial f(x^k) \quad \forall k \right\} \subseteq \partial f(x).
$$

Notice that when $f$ is convex, the above subdifferential coincides with the classical concept of convex subdifferential of $f$. Moreover, from the generalized Fermat’s rule, we know that if $x \in \mathbb{R}^m$ is a local minimizer of $f$, thus $0 \in \partial f(x)$.

Next, we briefly review the ALM, which is widely used in computer vision, image processing, statistical learning, see [1] for a comprehensive review. Let us consider the following minimization problem:

$$
\min_{x,y} f(x) + g(y) \quad \text{s.t.} \quad \mathcal{A}(x) + \mathcal{B}(y) = c,
$$

(5)

where $x, y \in \mathbb{R}^m$ are two variables, $\mathcal{A}, \mathcal{B} : \mathbb{R}^m \to \mathbb{R}^n$ are linear maps, and $c \in \mathbb{R}^n$ is a given vector. The augmented Lagrangian function for (5) is

$$
\mathcal{L}_\beta(x, y; w) := f(x) + g(y) - \langle w, \mathcal{A}(x) + \mathcal{B}(y) - c \rangle + \frac{\beta}{2} \|\mathcal{A}(x) + \mathcal{B}(y) - c\|^2,
$$

where $\beta > 0$ is a penalty parameter, and $w \in \mathbb{R}^n$ is the Lagrangian multiplier. Thus, the augmented Lagrangian method (ALM) to solving problem (5) can be summarized in the following iterative scheme:

$$
\left\{ \begin{array}{l}
{x^{k+1}, y^{k+1}} = \arg \min_{x,y} \{ \mathcal{L}_\beta(x, y; w^k) \}, \\
w^{k+1} = w^k - \tau \beta \left( \mathcal{A}(x^{k+1}) + \mathcal{B}(y^{k+1}) - c \right),
\end{array} \right.
$$

where $\tau \in \left( 0, 1 + \frac{\sqrt{5}}{2} \right)$. In most cases, the resulting subproblems admit closed-form solutions, which makes the ALM particularly efficient.

When $f$ and $g$ are convex functions, the ALM was originally proposed in [12] and the convergence was analyzed in [10, 11]. In fact, the convexity is the key property to characterize the convergence and the convergence rate of the ALM. But, when $f$ or $g$ is nonconvex, the convergence analysis of the ALM is much more challenging. This motivates the study of the ALM for nonconvex problems, see [13, 14, 23, 26] for illustrations. However, all the aforementioned results cannot be applied directly for solving (4). The difficulty is that the two variables share linear transformation, which makes the nonconvex problem more complex. In this paper, we attempt to overcome this difficulty and design a fast algorithm with convergence analysis.

At the end of this section, we review the following two lemmas which will help us express the closed-form solution for the proposed ALM algorithm.

**Lemma 2.3. ([2], Theorem 2.1) The minimization problem**

$$
\min_X \frac{1}{2} \|X - T\|_F^2 + \mu \|X\|_*
$$

has a closed-form solution which is given by

$$
X^* = U \text{Diag}(\text{shrink}_{2,1}(\sigma, \mu)) V^T,
$$

where $\text{shrink}_{2,1}$ is the shrinkage function and $U, V$ are the singular value decomposition of $X$.
where
\[
\text{shrink}_{2,1}(\sigma, \mu) = \text{sign}(\sigma) \max \left\{ 0, |\sigma| - \frac{1}{\mu} \right\}
\]
is the soft-thresholding defined in [5], and \( \Sigma = \text{Diag}(\sigma_1, \ldots, \sigma_m) \) is the singular value decomposition (SVD) of \( T \), i.e., \( T = U\Sigma V^T \).

**Lemma 2.4.** ([25], Section 3.2) For nonconvex functions
\[
\phi(x) = \sum_{i=1}^m \phi(x_i)
\]
with parameter \( \lambda > 0 \), the minimization problem
\[
\min_x \frac{1}{2} \|x - t\|^2 + \phi(x)
\]
has a closed-form solution which
\[
x^* = \text{shrink}_{2,\phi}(t, \lambda),
\]
where \( \text{shrink}_{2,\phi}(\cdot) \) is the associated shrinkage operators, see [25] for more details.

### 3. Proposed algorithm and convergence analysis.

To apply the ALM, we first introduce two auxiliary variables \( L \) and \( s \) and reformulate (4) as
\[
\min_{L, s, B, \gamma} \frac{1}{2} \sum_{i=1}^n (y_i - \langle X_i, B \rangle - \langle z_i, \gamma \rangle)^2 + \mu \|L\|_* + \phi(s)
\]
s.t. \( B = L, \gamma = s \).

The auxiliary variable \( L \) is used to replace \( B \) from the nonsmooth term \( \| \cdot \|_* \), and the variable \( s \) is to liberate \( \gamma \) from the nonsmooth term \( \phi(\cdot) \). Let \( \mathcal{L}_\beta(L, s, B, \gamma; W, w) \)
be the augmented Lagrangian function of (6) and it is defined as
\[
\mathcal{L}_\beta(L, s, B, \gamma; W, w) := \frac{1}{2} \sum_{i=1}^n (y_i - \langle X_i, B \rangle - \langle z_i, \gamma \rangle)^2 + \mu \|L\|_* + \phi(s)
\]
\[
-\langle W, B - L \rangle + \frac{\beta_1}{2} \|B - L\|_F^2 - \langle w, \gamma - s \rangle + \frac{\beta_2}{2} \|\gamma - s\|_2^2,
\]
in which \( W \in \mathbb{R}^{m \times q}, w \in \mathbb{R}^p \) are the Lagrangian multipliers and \( \beta_1, \beta_2 > 0 \) are the penalty parameters.

#### 3.1. Proposed algorithm.

Under the framework of ALM, the optimization problem of \( \mathcal{L}_\beta \) with respect to each variable can be solved by the following iterative scheme:
\[
\begin{align*}
(L_{k+1}^{s+1} & ) \quad \in \quad \arg \min_{L, s} \mathcal{L}_\beta(L, s, B^k, \gamma^k, W^k, w^k); \\
(B_{k+1}^{s+1} & ) \quad \in \quad \arg \min_{B, \gamma} \mathcal{L}_\beta(L^k, s^{k+1}, B, \gamma; W^k, w^k); \\
(W_{k+1}^{s+1} & ) \quad = \quad \left( \begin{array}{c}
W^k - \tau \beta_1 (B^{k+1} - L^{k+1}) \\
W^k - \tau \beta_2 (\gamma^{k+1} - s^{k+1})
\end{array} \right).
\end{align*}
\]

Now, we will show that each subproblem of the ALM either has a closed-form solution or can be solved by a fast solver.

The optimization subproblems of \( \mathcal{L}_\beta \) with respect to \( L \) and \( s \) are seperatable, thus they can be computed distributedly.
\( L^{k+1} = \arg \min_L \left\{ \mu \|L\|_s + \frac{\beta_1}{2} \left\| B^k - L - \frac{W^k}{\beta_1} \right\|_F^2 \right\}, \)

which can be solved by
\[
L^{k+1} = U \text{Diag} \left( \text{shrink}_{2,1}(\sigma_i, \frac{\mu}{\beta_1}) \right) V^T, \quad (7)
\]

where \( B^k - \frac{W^k}{\beta_1} = U \Sigma V^T \) and \( \Sigma = \text{Diag}(\sigma_1, \ldots, \sigma_m) \).

- **s-subproblem:**

\[
s^{k+1} = \arg \min_s \left\{ \phi(s) + \frac{\beta_2}{2} \left\| \gamma^k - s - \frac{w^k}{\beta_2} \right\|_2^2 \right\},
\]

and the solution is
\[
s^{k+1} = \text{shrink}_{2, \phi} \left( \gamma^{k+1} - \frac{w^k}{\beta_2}, \frac{1}{\beta_2} \right). \quad (8)
\]

The optimization subproblem of \( L_\beta \) associated with \( B \) and \( \gamma \) is

\[
\min_{B, \gamma} \frac{1}{2} \sum_{i=1}^n (y_i - \langle X_i, B \rangle - \langle z_i, \gamma \rangle)^2 + \frac{\beta_1}{2} \left\| B - L^{k+1} - \frac{W^k}{\beta_1} \right\|_F^2 + \frac{\beta_2}{2} \left\| \gamma - s^{k+1} - \frac{w^k}{\beta_2} \right\|_2^2.
\]

For convenience of description, we denote \( X := (\text{vec}(X_1), \ldots, \text{vec}(X_n))^T, y := (y_1, y_2, \ldots, y_n)^T, Z := (z_1, \ldots, z_n)^T \), then it can be transformed into

\[
\min_{B, \gamma} \frac{1}{2} \| y - X \text{vec}(B) - Z \gamma \|_2^2 + \frac{\beta_1}{2} \left\| B - L^{k+1} - \frac{W^k}{\beta_1} \right\|_F^2 + \frac{\beta_2}{2} \left\| \gamma - s^{k+1} - \frac{w^k}{\beta_2} \right\|_2^2.
\]

Note that the variables \( B \) and \( \gamma \) are not separable, thus we should compute at the same time. By grouping \( \text{vec}(B) \) and \( \gamma \), we obtain the following linear systems

\[
0 = - \left[ \begin{array}{c} X^T \\ Z^T \end{array} \right] \left( y - \left[ \begin{array}{c} X^T \\ Z^T \end{array} \right]^T \left[ \begin{array}{c} \text{vec}(B) \\ \gamma \end{array} \right] \right) + \left[ \begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right]^T \left( \left[ \begin{array}{c} \text{vec}(B^{k+1}) \\ \gamma^{k+1} \end{array} \right] - \left[ \begin{array}{c} \text{vec}(L^{k+1}) \\ s^{k+1} \end{array} \right] \right) - \left[ \begin{array}{c} \text{vec}(W^k) \\ w^k \end{array} \right].
\]

It is not hard to derive that

\[
\left[ \begin{array}{c} \text{vec}(B^{k+1}) \\ \gamma^{k+1} \end{array} \right] = \left[ \begin{array}{c} \beta_1 I + X^T X \\ \beta_2 I + Z^T Z \end{array} \right]^{-1} \left( \left[ \begin{array}{c} X^T \gamma \\ Z^T \gamma \end{array} \right] + \left[ \begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right]^T \left( \left[ \begin{array}{c} \text{vec}(B^{k+1}) \\ \gamma^{k+1} \end{array} \right] - \left[ \begin{array}{c} \text{vec}(L^{k+1}) \\ s^{k+1} \end{array} \right] \right) - \left[ \begin{array}{c} \text{vec}(W^k) \\ w^k \end{array} \right] \right).
\]

(9)

The coefficient matrices \( \beta_1 I + X^T X \) and \( \beta_2 I + Z^T Z \) are nonsingular whenever \( \beta_1, \beta_2 > 0 \), hence it can be computed via Cholesky decomposition or the conjugate gradient method.
The optimization subproblems of $L_\beta$ with respect to $W$ and $w$ are
\[
W^{k+1} = W^k - \tau_\beta_1 (B^{k+1} - L^{k+1}),
\]
\[
w^{k+1} = w^k - \tau_\beta_2 (\gamma^{k+1} - s^{k+1}).
\]
(10)

Finally, we present our proposed ALM for solving model (6) (equivalently (4)) in Algorithm 1.

---

Algorithm 1: ALM for solving (6)
Choose initial point $(B^0, \gamma^0, L^0, s^0; W^0, w^0)$.

while not converge do
  Step 1 Compute $(L^{k+1}, s^{k+1})$ according to (7) and (8);
  Step 2 Compute $(B^{k+1}, \gamma^{k+1})$ according to (9);
  Step 3 Update $(W^{k+1}, w^{k+1})$ via (10).
end while

3.2. Convergence analysis. In this section, we will prove the convergence of Algorithm 1 in detail. We first present the first-order optimality condition of (4) at the local minimizer $(B, \gamma)$, which is given by
\[
\begin{cases}
0 \in -\sum_{i=1}^n X_i (y_i - (X_i, B)) - (z_i, \gamma) + \mu \partial \bar{F}(B), \\
0 \in -\sum_{i=1}^n z_i (y_i - (X_i, B)) - \partial \phi(\gamma).
\end{cases}
\]
(11)

In this paper, we say that $(B^*, \gamma^*)$ is a stationary point of (4) if $(B^*, \gamma^*)$ satisfies (11) in place of $(B, \gamma)$. Then we present the first-order optimality conditions of the subproblems in Algorithm 1 as follows, which will be used repeatedly in our convergence analysis below.

\[
\begin{cases}
(a) 0 \in \partial \mu \| W^{k+1} \|_* + W^k - \beta_1 (B^k - L^{k+1}); \\
(b) 0 \in \partial \phi (s^{k+1}) + w^k - \beta_2 (\gamma^k - s^{k+1}); \\
(c) 0 = -\sum_{i=1}^n X_i (y_i - (X_i, B^{k+1}) - (z_i, \gamma^{k+1})) - W^k + \beta_1 (B^{k+1} - L^{k+1}); \\
(d) 0 = -\sum_{i=1}^n z_i (y_i - (X_i, B^{k+1}) - (z_i, \gamma^{k+1})) - w^k + \beta_2 (\gamma^{k+1} - s^{k+1}); \\
(e) W^{k+1} - W^k = -\tau_\beta_1 (B^{k+1} - L^{k+1}); \\
f) w^{k+1} - w^k = -\tau_\beta_2 (\gamma^{k+1} - s^{k+1}).
\end{cases}
\]
(12)

Our convergence is largely based on the following potential function:
\[
\Theta_{\beta, \eta}(L, s, B, \gamma; \lambda) := L_\beta (L, s, B, \gamma; \Lambda, \lambda) + \eta (\| B - L \|_F^2 + \beta_2 \| \gamma - s \|_2^2),
\]
where $\eta$ is a tuning parameter, and
\[
\eta := \max \left\{ 1 - \tau, \frac{(\tau - 1) \tau^2}{1 + \tau - \tau^2} \right\}.
\]

Lemma 3.1. Suppose that $\{L^k, s^k, B^k, \gamma^k; W^k, w^k\}$ is a sequence generated by Algorithm 1. Then we have
\[
\Theta_{\beta, \eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^{k+1}, w^{k+1}) - \Theta_{\beta, \eta}(L^k, s^k, B^k, \gamma^k; W^k, w^k) \leq \left( \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2} \right\} \cdot \frac{2 \lambda_{\max}^2}{\beta_1} \right) \cdot \frac{\lambda_{\max}^2 \beta_2}{2} \| B^{k+1} - B^k \|_F^2 \\
+ \left( \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2} \right\} \cdot \left( \frac{2 \lambda_{\max}^2 \beta_1}{\beta_2^2} + \frac{\lambda_{\max}^2 \beta_2}{2} \right) \right) \| \gamma^{k+1} - \gamma^k \|_2^2.
\]
Moreover, if $\beta_1, \beta_2$ satisfy some mild conditions, the it is easy to conclude that the sequence $\{\Theta_{\beta, \eta}(L^k, s^k, B^k, \gamma^k; W^k, w^k)\}_{k=1}^\infty$ is decreasing.

Proof. First, we notice that
\begin{align*}
\Theta_{\beta, \eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^{k+1}, w^{k+1}) \\
- \Theta_{\beta, \eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^k, w^k) \\
= \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^{k+1}, w^{k+1}) \\
- \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^k, w^k) \\
= -\langle \Lambda^{k+1} - \Lambda^k, B_k^{k+1} - L^{k+1} \rangle - (\lambda^{k+1} - \lambda^k, \gamma^{k+1} - s^{k+1}) \\
= \frac{1}{\tau^2} ||W^{k+1} - W^k||_F^2 + \frac{1}{\tau^2} ||w^{k+1} - w^k||_2^2, \tag{13}
\end{align*}
where the third equality follows from (e) and (f). We then analyze the upper bound for $\frac{1}{\tau^2} ||W^{k+1} - W^k||_F^2$, and similar result can be derived for $\frac{1}{\tau^2} ||w^{k+1} - w^k||_2^2$.

From (c), we have
\begin{align*}
0 &= -\sum_{i=1}^n X_i(y_i - \langle X_i, B^{k+1} \rangle - \langle z_i, \gamma^{k+1} \rangle) - W^k + \beta_i(B^{k+1} - L^{k+1}) \\
&= -\sum_{i=1}^n X_i(y_i - \langle X_i, B^{k+1} \rangle - \langle z_i, \gamma^{k+1} \rangle) - W^k - \frac{1}{\tau}(W^{k+1} - W^k),
\end{align*}
where the last equality follows from (e), and it can be transformed to
\begin{align*}
W^{k+1} &= -\tau \sum_{i=1}^n X_i(y_i - \langle X_i, B^{k+1} \rangle - \langle z_i, \gamma^{k+1} \rangle) + (1 - \tau)W^k.
\end{align*}

Hence, for $k > 1$,
\begin{align*}
W^{k+1} - W^k &= -\tau \sum_{i=1}^n X_i(y_i - \langle X_i, B^{k+1} \rangle - \langle z_i, \gamma^{k+1} \rangle) + (1 - \tau)W^k \\
&\quad + \tau \sum_{i=1}^n X_i(y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle) - (1 - \tau)W^{k-1} \\
&= \tau \sum_{i=1}^n X_i((X_i, B^{k+1} - B^k) + \langle z_i, \gamma^k - \gamma^{k+1} \rangle) + (1 - \tau)(W^k - W^{k-1}). \tag{14}
\end{align*}

We now consider two separate cases: $0 < \tau \leq 1$ and $1 < \tau < \frac{1 + \sqrt{5}}{2}$.

- For $0 < \tau \leq 1$, it follows that
\begin{align*}
||\Lambda^{k+1} - \Lambda^k||_F^2 &\leq ||\tau \sum_{i=1}^n X_i((X_i, B^{k+1} - B^k) + \langle z_i, \gamma^k - \gamma^{k+1} \rangle) + (1 - \tau)(W^k - W^{k-1})||_F^2 \\
&\leq \tau \sum_{i=1}^n X_i((X_i, B^{k+1} - B^k) + \langle z_i, \gamma^k - \gamma^{k+1} \rangle)||_F^2 + (1 - \tau)||W^k - W^{k-1}||_F^2 \\
&\leq 2\tau \sum_{i=1}^n X_i((X_i, B^{k+1} - B^k)||_F^2 + 2\tau \sum_{i=1}^n X_i(z_i, \gamma^k) \\
&\quad - \gamma^{k+1})||_F^2 + (1 - \tau)||W^k - W^{k-1}||_F^2 \\
&\leq 2\tau(\lambda^X_{\max})^2||B^{k+1} - B^k||_F^2 + 2\tau\lambda^X_{\max}\lambda^Z_{\max}||\gamma^{k+1} \\
&\quad - \gamma^k||_F^2 + (1 - \tau)||W^k - W^{k-1}||_F^2,
\end{align*}
where $\lambda^X_{\max}$ is the largest eigenvalue of $XX^T$, and $\lambda^Z_{\max}$ is the largest eigenvalue of $ZZ^T$. The first inequality comes from the convexity of $||\cdot||_F^2$, the second
inequality follows from basic inequality properties, i.e., $(u + v)^2 \leq 2u^2 + 2v^2$, the third inequality stems from the property of $\lambda^X_{\max}$ and $\lambda^Z_{\max}$. We then add $-(1 - \tau)\|W^{k+1} - W^k\|^2_F$ to both sides of the above inequality and simplify the resulting inequality to get
\[
\|W^{k+1} - W^k\|^2_F \leq 2(\lambda^X_{\max})^2 \|B^{k+1} - B^k\|^2_F + 2\lambda^X_{\max} \lambda^Z_{\max} \|\gamma^{k+1} - \gamma^k\|^2_F \\
+ \frac{1 - \tau}{\tau} \left( \|W^k - W^{k-1}\|^2_F - \|W^{k+1} - W^k\|^2_F \right) \\
= 2(\lambda^X_{\max})^2 \|B^{k+1} - B^k\|^2_F + 2\lambda^X_{\max} \lambda^Z_{\max} \|\gamma^{k+1} - \gamma^k\|^2_F \\
+ (1 - \tau)\beta_1^2 \left( \|B^k - L^k\|^2_F - \|B^{k+1} - L^{k+1}\|^2_F \right),
\]
where the last equality follows from (e) in (12).

- For $1 < \tau < \frac{1 + \sqrt{5}}{2}$, dividing $\tau$ from both sides of (14), we have
\[
\frac{1}{\tau} (\Lambda^{k+1} - \Lambda^k) = \sum_{i=1}^n X_i (\langle X_i, B^{k+1} - B^k \rangle + \langle z_i, \gamma^k - \gamma^{k+1} \rangle) \\
+ \frac{1 - \tau}{\tau} (W^k - W^{k-1}) \\
= \frac{1}{\tau} \sum_{i=1}^n X_i (\langle X_i, B^{k+1} - B^k \rangle + \langle z_i, \gamma^k - \gamma^{k+1} \rangle) + \left( 1 - \frac{1}{\tau} \right) (W^{k-1} - W^k).
\]
This together with $0 < \frac{1}{\tau} < 1$ and the convexity of $\| \cdot \|^2_F$, implies that
\[
\left\| \frac{1}{\tau} (W^{k+1} - W^k) \right\|^2_F \\
\leq \frac{1}{\tau} \left( \frac{1}{\tau} \right) \sum_{i=1}^n X_i (\langle X_i, B^{k+1} - B^k \rangle + \langle z_i, \gamma^k - \gamma^{k+1} \rangle) \left\| W^k - W^{k-1} \right\|^2_F \\
+ \left( 1 - \frac{1}{\tau} \right) \left\| W^{k-1} - W^k \right\|^2_F \\
\leq 2\tau(\lambda^X_{\max})^2 \|B^{k+1} - B^k\|^2_F + 2\tau\lambda^X_{\max} \lambda^Z_{\max} \|\gamma^{k+1} - \gamma^k\|^2_F \\
+ \left( 1 - \frac{1}{\tau} \right) \left\| W^k - W^{k-1} \right\|^2_F.
\]
Hence,
\[
\|W^{k+1} - W^k\|^2_F \\
\leq 2\tau^3(\lambda^X_{\max})^2 \|B^{k+1} - B^k\|^2_F + 2\tau^3\lambda^X_{\max} \lambda^Z_{\max} \|\gamma^{k+1} - \gamma^k\|^2_F \\
- \|W^{k+1} - W^k\|^2_F.
\]
Then, adding $-(\tau^2 - \tau)\|W^{k+1} - W^k\|^2_F$ to both sides of the above inequality, simplifying the resulting and using the fact $1 + \tau - \tau^2 > 0$ for $1 < \tau < \frac{1 + \sqrt{5}}{2}$, we see that
\[
\|W^{k+1} - W^k\|^2_F \\
\leq \frac{2\tau^3(\lambda^X_{\max})^2}{1 + \tau - \tau^2} \|B^{k+1} - B^k\|^2_F + \frac{2\tau^3\lambda^X_{\max} \lambda^Z_{\max}}{1 + \tau - \tau^2} \|\gamma^{k+1} - \gamma^k\|^2_F \\
+ \frac{\tau^2 - \tau}{1 + \tau - \tau^2} \left( \|W^k - W^{k-1}\|^2_F - \|W^{k+1} - W^k\|^2_F \right).
\]
Thus, for $0 < \tau < \frac{1+\sqrt{5}}{2}$, combining (15), (16) and recalling the definition of $\eta$ in $\Theta_{\beta,\tau}$, we have

$$
\frac{1}{\tau \beta_1} \|W^{k+1} - W^k\|_F^2 \leq \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \frac{2\lambda X_{\max}}{\beta_1} \|B^{k+1} - B^k\|_F^2 \\
+ \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \frac{2\lambda Z_{\max}}{\beta_1} \|\gamma^{k+1} - \gamma^k\|_2^2 \\
+ \eta \beta_1 \left( \|B^k - L^k\|_F^2 - \|B^{k+1} - L^{k+1}\|_F^2 \right). \tag{17}
$$

Similarly, from (d), we have

$$w^{k+1} - w^k = \tau \sum_{i=1}^n z_i (\langle X_i, B^k - B^{k+1} \rangle + \langle z_i^T, \gamma^k - \gamma^{k+1} \rangle) + (1 - \tau) (w^k - w^{k-1}).$$

Then, we can also derive

$$
\frac{1}{\tau \beta_2} \|\lambda^{k+1} - \lambda^k\|_2^2 \leq \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \frac{2\lambda X_{\max}}{\beta_2} \|B^{k+1} - B^k\|_F^2 \\
+ \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \frac{2\lambda Z_{\max}}{\beta_2} \|\gamma^{k+1} - \gamma^k\|_2^2 \\
+ \eta \beta_2 \left( \|\gamma^k - s^k\|_2^2 - \|\gamma^{k+1} - s^{k+1}\|_2^2 \right). \tag{18}
$$

Next, note that the function $\gamma \leftarrow \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma; W^k, w^k)$ is strongly convex with modulus at least $\lambda_{\max}^Z + \beta_2$. Using this fact and the definition of $\gamma^{k+1}$, we see that

$$\Theta_{\beta,\eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^k, w^k) - \Theta_{\beta,\eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^k; W^k, w^k) = \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^{k+1}; W^k, w^k) - \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^k; W^k, w^k) \\
+ \eta \beta_2 \left( \|\gamma^{k+1} - s^{k+1}\|_2^2 - \|\gamma^k - s^k\|_2^2 \right) \\
\leq \eta \beta_2 \left( \|\gamma^{k+1} - s^{k+1}\|_2^2 - \|\gamma^k - s^k\|_2^2 \right) - \frac{\lambda_{\max}^Z + \beta_2}{2} \|\gamma^{k+1} - \gamma^k\|_2^2. \tag{19}
$$

Then, $B \leftarrow \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B, \gamma^k; W^k, w^k)$ is also strongly convex with modulus at least $\lambda_{\min}^X + \beta_1$, and we have

$$\Theta_{\beta,\eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^k; W^k, w^k) - \Theta_{\beta,\eta}(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^k; W^k, w^k) = \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^k; W^k, w^k) - \mathcal{L}_\beta(L^{k+1}, s^{k+1}, B^{k+1}, \gamma^k; W^k, w^k) \\
+ \eta \beta_1 \left( \|B^{k+1} - L^{k+1}\|_F^2 - \|B^k - L^{k+1}\|_F^2 \right) \\
\leq \eta \beta_1 \left( \|B^{k+1} - L^{k+1}\|_F^2 - \|B^k - L^{k+1}\|_F^2 \right) - \frac{\lambda_{\min}^X + \beta_1}{2} \|B^{k+1} - B^k\|_F^2. \tag{20}
$$
Moreover, because $s^{k+1}$ is a minimizer of $s$-subproblem, we have

$$\Theta_{\beta_1}(L^{k+1}, s^{k+1}, B^k, \gamma^k; W, w^k) = \Theta_{\beta_1}(L^k, s^k, B^k, \gamma^k; W, w^k)$$

Thus, if

$$\max\left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \left(2\frac{(\lambda_{\max}^X)^2}{\beta_1} + 2\frac{\lambda_{\max}^X \lambda_{\max}^Z}{\beta_2} \right) - \frac{\lambda_{\min}^X + \beta_1}{2} \leq 0$$

and

$$\max\left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \left(2\frac{(\lambda_{\max}^Z)^2}{\beta_2} + 2\frac{\lambda_{\max}^X \lambda_{\max}^Z}{\beta_1} \right) - \frac{\lambda_{\min}^Z + \beta_2}{2} \leq 0,$$

then the sequence $\{\Theta_{\beta_1}(L^k, s^k, B^k, \gamma^k; W, w^k)\}_{k=1}^{\infty}$ is decreasing. This completes the proof.

Lemma 3.1 has proved that the sequence $\{\Theta_{\beta_1}(L^k, s^k, B^k, \gamma^k; W, w^k)\}_{k=1}^{\infty}$ generated by Algorithm 1 is decreasing under some mild conditions. Moreover, it is not hard to check the existence of the solution. For notational simplicity, from now on, we denote the conditions as $\beta_1 \geq \beta_1^*, \beta_2 \geq \beta_2^*$. We next illustrate that the corresponding sequence $\{(L^k, s^k, B^k, \gamma^k; W, w^k)\}_{k=1}^{\infty}$ is also bounded.

**Lemma 3.2.** Suppose that $\beta_1 \geq \beta_1^*, \beta_2 \geq \beta_2^*$ and $\{(L^k, s^k, B^k, \gamma^k; W, w^k)\}$ is a sequence generated by Algorithm 1. Then the sequence $\{(L^k, s^k, B^k, \gamma^k; W, w^k)\}_{k=1}^{\infty}$ is bounded.

**Proof.** From Lemma 3.1, we know the sequence $\{\Theta_{\beta_1}(L^k, s^k, B^k, \gamma^k; W, w^k)\}_{k=1}^{\infty}$ generated by Algorithm 1 is decreasing. Suppose $\{\Theta_{\beta_1}(L^k, s^k, B^k, \gamma^k; W, w^k)\}_{k=1}^{\infty}$
has a upper bound, say \( \Theta_{\tau, \beta}(L^1, s^1, B^1, \gamma^1; W^1, w^1) \). Thus, for \( k \geq 1 \), we have

\[
\Theta_{\beta, \gamma}(L^k, s^k, B^k, \gamma^k; W^k, w^k) \geq \Theta_{\beta, \gamma}(L^{k}, s^{k}, B^{k}, \gamma^{k}; W^{k}, w^{k})
\]

\[
= \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + \mu \|L^k\|_\ast + \phi(s^k)
\]

\[
- \langle W^k, B^k - L^k \rangle + \frac{\beta_1}{2} \|B^k - L^k\|_F^2
\]

\[
- \langle \lambda^k, \gamma^k - s^k \rangle + \frac{\beta_2}{2} \|\gamma^k - s^k\|_2^2 + \eta(\beta_1 \|B^k - L^k\|_F^2 + \beta_2 \|\gamma^k - s^k\|_2^2)
\]

\[
= \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + \mu \|L^k\|_\ast + \phi(s^k) + \frac{\beta_1}{2} \|B^k - L^k - \frac{W^k}{\beta_1}\|_F^2
\]

\[
+ \frac{\beta_2}{2} \|\gamma^k - s^k - \frac{w^k}{\beta_2}\|_2^2 - \frac{1}{2\beta_1\beta_2} \|\Lambda^k\|_F^2 - \frac{1}{2\beta_1\beta_2} \|\lambda^k\|_2^2
\]

\[
+ \eta(\beta_1 \|B^k - L^k\|_F^2 + \beta_2 \|\gamma^k - s^k\|_2^2).
\]

(24)

We next derive the upper bound for \( \|W^k\|_F^2 \). We start by substituting (e) into (c) to obtain

\[
0 = \sum_{i=1}^{n} X_i(y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle) + W^{k-1} + \frac{1}{\tau}(W^k - W^{k-1}),
\]

which yields

\[
\tau W^k = \tau \sum_{i=1}^{n} X_i(y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle) + (1 - \tau)(W^k - W^{k-1}).
\]

(25)

We now consider two different cases:

- For \( 0 < \tau \leq 1 \), it follows from the convexity of \( \| \cdot \|_F^2 \) and (25) that

\[
\|\tau W^k\|_F^2 \leq \tau \sum_{i=1}^{n} X_i(y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + (1 - \tau) \|W^k - W^{k-1}\|_F^2.
\]

\[
\leq \tau \lambda^\ast \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + (1 - \tau) \|W^k - W^{k-1}\|_F^2.
\]

\[
= \tau \lambda^\ast \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + (1 - \tau) \|W^k - W^{k-1}\|_F^2.
\]

\[
\leq 2\tau \lambda^\ast \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + (1 - \tau) \|W^k - W^{k-1}\|_F^2.
\]

where the equality follows from (e) and the last inequality comes from the basic inequality property. Then, we have

\[
\|W^k\|_F^2 \leq \frac{2\lambda^\ast}{\tau} \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + (1 - \tau) \|W^k - W^{k-1}\|_F^2.
\]

(26)
For $1 < \tau < \frac{1 + \sqrt{5}}{2}$, dividing $\tau$ from both sides of (25), we obtain

$$\Lambda^k = \frac{1}{\tau} \sum_{i=1}^{n} X_i(y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle) + \left(1 - \frac{1}{\tau}\right) (W^k - W^{k-1}).$$

Then, since $0 < \frac{1}{\tau} < 1$, using the convexity of $\| \cdot \|_F$ and (e), we have

$$\|W^k\|_F^2 \leq \frac{1}{\tau} \sum_{i=1}^{n} X_i(y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + \left(1 - \frac{1}{\tau}\right) \|W^k - W^{k-1}\|_F^2$$

$$\leq \tau \lambda^X \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2 + \tau \beta_1^2 \|B^k - L^k\|_F^2.$$  \hspace{1cm} (27)

Thus, combining (26) and (27), we have

$$\|W^k\|_F^2 \leq \max\{1/\tau, \tau\} \cdot 2\lambda^X \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2$$

$$+ \max\{1 - \tau, (\tau - 1)\beta_2\beta_1^2 \|B^k - L^k\|_F^2,$$

then,

$$- \frac{1}{2\beta_1} \|W^k\|_F^2$$

$$\geq - \max\{1/\tau, \tau\} \lambda^X \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2$$

$$- \max\{1 - \tau, (\tau - 1)\beta_2\beta_1\} \|B^k - L^k\|_F^2.$$  \hspace{1cm} (28)

Similarly, for $\|w^k\|_2^2$, we can also derive that

$$- \frac{1}{2\beta_2} \|w^k\|_2^2$$

$$\geq - \max\{1/\tau, \tau\} \lambda^Z \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2$$

$$- \max\{1 - \tau, (\tau - 1)\beta_2\beta_1\} \|\gamma^k - s^k\|_F^2.$$  \hspace{1cm} (29)

Putting (28) and (29) into (24), we have

$$\Theta_{\beta, \eta}(L^1, s^1, B^1, \gamma^1; W^1, w^1) \geq \mu \|L^k\|_* + \phi(s^k)$$

$$+ \left(1 - \max\{1/\tau, \tau\} \cdot \left(\frac{\lambda^X}{\beta_1} + \frac{\lambda^Z}{\beta_2}\right)\right) \sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2$$

$$+ \frac{\beta_1}{2} \|B^k - L^k\|_F^2 + \frac{\beta_2}{2} \|\gamma^k - s^k\|_F^2$$

$$+ \left(\eta - \max\{1 - \tau, (\tau - 1)\beta_2\beta_1\}\right) \left(\beta_1 \|B^k - L^k\|_F^2 + \beta_2 \||\gamma^k - s^k\|_F^2\right).$$  \hspace{1cm} (30)
With (30) established, we are now ready to prove the boundedness of the sequence. We start with the observation that for $0 < \tau < \frac{1 + \sqrt{2}}{2}$, $\beta_1 \geq \beta_1$, $\beta_2 \geq \beta_2$, and the definition of $\eta$, we always have
\[
1 - \max\{1/\tau, \tau\} \cdot \left(\frac{\lambda_{\max}^X}{\beta_1} + \frac{\lambda_{\max}^Z}{\beta_2}\right) \geq 0,
\]
and
\[
\eta \geq \frac{1 - \tau, (\tau - 1)\tau}{2} \geq 0.
\]
This, together with the fact that $\|L^k\|_*$ and $\phi(s^k)$ is nonnegative, illustrates these items are all bounded, including $\sum_{i=1}^{n} (y_i - \langle X_i, B^k \rangle - \langle z_i, \gamma^k \rangle)^2$, $\|B^k - L^k - \frac{w^k}{\tau}\|^2_F$, $\|\gamma^k - s^k - \frac{w^k}{\tau}\|^2_2$, $\|\gamma^k - s^k\|_2^2$, and $\|\gamma^k - s^k\|_2^2$. Moreover,
\[
\mu \|L^k\|_* + \phi(s^k) \leq \Theta_{\beta, \eta}(L^1, s^1, B^1, \gamma^1; W^1, w^1),
\]
from which we see immediately that $\{L^k\}$ and $\{s^k\}$ are bounded, and then $\{B^k\}$ and $\{\gamma^k\}$ are also bounded. The boundedness of $\{W^k\}$ and $\{w^k\}$ now follow from (28) and (29). Now, we conclude that the sequence $\{(L^k, s^k, B^k, \gamma^k; W^k, w^k)\}_{k=1}^{\infty}$ is bounded and the proof is completed.

Before stating the convergence result, we also need the following lemma.

**Lemma 3.3.** Suppose that $\beta_1 \geq \beta_1$, $\beta_2 \geq \beta_2$ and $\{(L^k, s^k, B^k, \gamma^k; W^k, w^k)\}$ is a sequence generated by Algorithm 1. Then we have
\[
\lim_{k \to \infty} \|L^{k+1} - L^k\|_F + \|s^{k+1} - s^k\|_2 + \|B^{k+1} - B^k\|_F + \|\gamma^{k+1} - \gamma^k\|_2 + \|w^{k+1} - w^k\|_2 = 0.
\]

**Proof.** Suppose that $\{(L^*, s^*, B^*, \gamma^*; W^*, w^*)\}$ is a cluster point of a sequence $\{(L^k, s^k, B^k, \gamma^k; W^k, w^k)\}$ generated by Algorithm 1, and $\{(L^{k_i}, s^{k_i}, B^{k_i}, \gamma^{k_i}; W^{k_i}, w^{k_i})\}$ is a convergent subsequence satisfies
\[
\lim_{i \to \infty} (L^{k_i}, s^{k_i}, B^{k_i}, \gamma^{k_i}; W^{k_i}, w^{k_i}) = (L^*, s^*, B^*, \gamma^*; W^*, w^*).
\]

Summing (23) from $k = 1$ to $k = k_i - 1$, we have
\[
\Theta_{\beta, \eta}(L^{k_i}, s^{k_i}, B^{k_i}, \gamma^{k_i}; W^{k_i}, w^{k_i}) - \Theta_{\beta, \eta}(L^1, s^1, B^1, \gamma^1; W^1, w^1) \leq -c_1 \sum_{k=1}^{k_i-1} \|B^{k+1} - B^k\|^2_F - c_2 \sum_{k=1}^{k_i-1} \|\gamma^{k+1} - \gamma^k\|^2_2,
\]
where
\[
c_1 := -\max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \left(2 \frac{(\lambda_{\max}^X)^2}{\beta_1} + 2 \frac{\lambda_{\max}^X}{\beta_1} + \frac{\lambda_{\max}^Z}{\beta_2}\right) - \frac{\lambda_{\min}^X + \beta_1}{2}
\]
and
\[
c_2 := -\max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \left(2 \frac{(\lambda_{\max}^Z)^2}{\beta_2} + 2 \frac{\lambda_{\max}^Z}{\beta_2} + \frac{\lambda_{\min}^Y + \beta_2}{2}\right).
\]
Passing to the limit and rearranging terms in the resulting relation, we obtain
\[
c_1 \sum_{k=1}^{\infty} \|B^{k+1} - B^k\|^2_F + c_2 \sum_{k=1}^{\infty} \|\gamma^{k+1} - \gamma^k\|^2_2 \leq \Theta_{\beta, \eta}(L^1, s^1, B^1, \gamma^1; W^1, w^1) - \Theta_{\beta, \eta}(L^*, s^*, B^*, \gamma^*; W^*, w^*) < \infty.
\]
This together with $c_1, c_2 \geq 0$ derives
\[
\sum_{k=1}^{\infty} \|B^{k+1} - B^k\|_F^2 < \infty, \sum_{k=1}^{\infty} \|\gamma^{k+1} - \gamma^k\|_2^2,
\]
which implies
\[
B^{k+1} - B^k \to 0, \ \gamma^{k+1} - \gamma^k \to 0.
\]
Next, summing both sides of (17) from $k = 1$ to $k = k_i$ and taking limits, we have
\[
\frac{1}{\tau \beta} \sum_{k=1}^{\infty} \|W^{k+1} - W^k\|_F^2 \leq \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot 2(\lambda\max^2) \sum_{k=1}^{\infty} \|B^{k+1} - B^k\|_F^2 + \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot 2(\lambda\max^2) \sum_{k=1}^{\infty} \|\gamma^{k+1} - \gamma^k\|_2^2 \]
\[+ \eta \beta \left( \|B^1 - L^1\|_F^2 - \lim \inf_{k \to \infty} \|B^{k+1} - L^{k+1}\|_F^2 \right) < \infty,
\]
from which we conclude that
\[
W^{k+1} - W^k \to 0.
\]
Similarly, we can also get
\[
w^{k+1} - w^k \to 0.
\]
Finally, by combing (31), (32), (33), and (c-d) in (12), we have
\[
L^{k+1} - L^k \to 0, \ s^{k+1} - s^k \to 0.
\]
Therefore, we have completed the proof. \qed

Based on the above three lemmas, we are now ready to prove our convergence result for Algorithm 1, which also characterizes the cluster point of the sequence.

**Theorem 3.4.** Suppose that $\beta_1 \geq \beta_1$, $\beta_2 \geq \beta_2$ and \{(\$L^k, s^k, B^k, \gamma^k; W^k, w^k$)\} is a sequence generated by Algorithm 1. Then any cluster point \{(\$L^*, s^*, B^*, \gamma^*; W^*, w^*$)\} of the sequence is a stationary point of (4).

**Proof.** Indeed, $\Theta_{\tau, \beta}(\cdot)$ is lower semicontinuous, we have
\[
\liminf_{i \to \infty} \Theta_{\beta_1, \beta_2}(L^{k_i+1}, s^{k_i+1}, B^{k_i}, \gamma^k; W^{k_i}, w^{k_i})
\geq \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle)^2 + \mu\|L^*\|_F + \phi(s^*) - \langle W^*, B^* - L^* \rangle
\]
\[+ \beta \|B^* - L^*\|_F^2 - \|w^*, \gamma^* - s^*\|^2 / 2 + \beta \|\gamma^* - s^*\|_2^2 \]
\[+ \eta \beta_1 \|B^* - L^*\|_F^2 + \beta_2 \|\gamma^* - s^*\|_2^2. \]
(34)

On the other hand, from the definition of $L^{k_i+1}$ and $s^{k_i+1}$, we obtain
\[
\Theta_{\beta_1, \beta_2}(L^{k_i+1}, s^{k_i+1}, B^{k_i}, \gamma^k; W^{k_i}, w^{k_i})
\leq \Theta_{\beta_1, \beta_2}(L^*, s^*, B^{k_i}, \gamma^k; W^{k_i}, w^{k_i}) + \eta (\beta_1 \|B^{k_i} - L^*\|_F - \beta_2 \|\gamma^k - s^*\|_2^2).
\]
Specifically, we initialize with a small step-size. In practice, we can apply an incremental process.

Remark. From Theorem 3.4, the penalty parameter $\beta_1, \beta_2$ should be chosen to be $\beta_1 \geq \bar{\beta}_1, \beta_2 \geq \bar{\beta}_2$. Note that it is not hard to check the existence. In fact, $\beta_1, \beta_2$ are sufficiently large, the ALM algorithm will be very slow, because $\beta_1, \beta_2$ are associated with step-size. In practice, we can apply an incremental process. Specifically, we initialize with a small $\beta_1, \beta_2$ (less than $\bar{\beta}_1, \bar{\beta}_2$), and then increase the $\beta_1, \beta_2$ by a constant ratio until $\beta_1 \geq \bar{\beta}_1, \beta_2 \geq \bar{\beta}_2$. Clearly, after at most finitely steps, the penalty parameters $\beta_1, \beta_2$ get above the threshold $\bar{\beta}_1, \bar{\beta}_2$ and the convergence of the resulting algorithm is guaranteed.

Taking limits on both sides of this inequality, we can derive

$$
\liminf_{i \to \infty} \Theta_{\beta, \eta}(L^{k_i+1}, s^{k_i+1}, B^{k_i}, \gamma^{k_i}; W^{k_i}, w^{k_i})
\leq \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle)^2 + \mu \|L^*\|_* + \phi(s^*) - \langle W^*, B^* - L^* \rangle
+ \frac{\beta}{2} \|B^* - L^*\|_F^2 - \langle w^*, \gamma^* - s^* \rangle + \frac{\beta}{2} \|\gamma^* - s^*\|_2^2
+ \eta(\beta_1 \|B^* - L^*\|_F^2 + \beta_2 \|\gamma^* - s^*\|_2^2).
$$

Then, together with (34) and (35), we see that

$$
\liminf_{i \to \infty} \Theta_{\beta, \eta}(L^{k_i+1}, s^{k_i+1}, B^{k_i}, \gamma^{k_i}; W^{k_i}, w^{k_i})
= \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle)^2 + \mu \|L^*\|_* + \phi(s^*) - \langle W^*, B^* - L^* \rangle
+ \frac{\beta}{2} \|B^* - L^*\|_F^2 - \langle w^*, \gamma^* - s^* \rangle + \frac{\beta}{2} \|\gamma^* - s^*\|_2^2
+ \eta(\beta_1 \|B^* - L^*\|_F^2 + \beta_2 \|\gamma^* - s^*\|_2^2).
$$

It follows from $L^{k_i+1} - L^k \to 0$, $s^{k_i+1} - s^k \to 0$ in Lemma 3.3 that

$$
\lim_{i \to \infty} \|L^{k_i+1}\|_* = \|L^*\|_*, \lim_{i \to \infty} \phi(s^{k_i+1}) = \phi(s^*).
$$

Thus, passing the limit in (a-f) in (12) along \{(L^{k_i}, s^{k_i}, B^{k_i}, \gamma^{k_i}; W^{k_i}, w^{k_i})\}, and invoking statement (36), we see that

$$
\begin{align*}
0 & \in \partial \mu \|L^*\|_* + W^* - \beta_1 (B^* - L^*); \\
0 & \in \partial \phi(s^*) + w^* - \beta_2 (\gamma^* - s^*); \\
0 & = -\sum_{i=1}^{n} X_i(y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle) - W^* + \beta_1 (B^* - L^*); \\
0 & = -\sum_{i=1}^{n} z_i(y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle) - w^* + \beta_2 (\gamma^* - s^*); \\
B^* & = L^*; \\
\gamma^* & = s^*.
\end{align*}
$$

Rearranging terms in the above relations, we immediately have

$$
\begin{align*}
0 & \in -\sum_{i=1}^{n} X_i(y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle) + \mu \|B\|_*; \\
0 & \in -\sum_{i=1}^{n} z_i(y_i - \langle X_i, B^* \rangle - \langle z_i, \gamma^* \rangle) + \partial \phi(s^*).
\end{align*}
$$

This shows that $(L^*, s^*, B^*, \gamma^*; W^*, w^*)$ gives a stationary point of (4). This completes the proof. \qed

Remark. From Theorem 3.4, the penalty parameter $\beta_1, \beta_2$ should be chosen to be $\beta_1 \geq \bar{\beta}_1, \beta_2 \geq \bar{\beta}_2$. Note that it is not hard to check the existence. In fact, $\beta_1, \beta_2$ are sufficiently large, the ALM algorithm will be very slow, because $\beta_1, \beta_2$ are associated with step-size. In practice, we can apply an incremental process. Specifically, we initialize with a small $\beta_1, \beta_2$ (less than $\bar{\beta}_1, \bar{\beta}_2$), and then increase the $\beta_1, \beta_2$ by a constant ratio until $\beta_1 \geq \bar{\beta}_1, \beta_2 \geq \bar{\beta}_2$. Clearly, after at most finitely steps, the penalty parameters $\beta_1, \beta_2$ get above the threshold $\bar{\beta}_1, \bar{\beta}_2$ and the convergence of the resulting algorithm is guaranteed.
4. **Numerical experiments.** In this section, we will conduct some experiments on synthetic and real-world data to demonstrate the superiority of our proposed NonMM (4) over LMM (2) and LSMM (3). For the sake of fairness, we run 100 times to illustrate the performance of above algorithms. All the experiments are performed using MATLAB (R2018b) on a laptop computer with an Intel Core i7 CPU with 1.4 GHz and 16 GB of memory.

For the LMM, the results are generated from the source codes released by their authors, and the parameters are all set as default. For the LSMM, we implement it by ourselves. For our proposed NonMM, we choose $\phi$ to be the SCAD with $a = 3.7$ for the nonconvex regularization. As is well known to us, the accuracy of the solution depends on the latent parameters $\lambda$. We empirically set

$$\mu = \alpha_1 \| X^T y \|_{\infty}, \quad \lambda = \alpha_2 \mu,$$

where $0 < \alpha_1 < 1$ and $\alpha_2 > 0$. In addition, for all test algorithms, the maximum iteration number is set as 50000, and the stopping criterion is set to be $\text{Tol} < 10^{-4}$, which is defined as

$$\text{Tol} = \max \left\{ \frac{\| B^{k+1} - B^k \|_F}{\| B^k \|_F + 1}, \frac{\| \gamma^{k+1} - \gamma^k \|}{\| \gamma^k \| + 1} \right\} < 10^{-4}.$$

4.1. **Comparison results for synthetic data.** In this subsection, we test synthetic data to compare the LMM, LSMM and NonMM. According to [31], we first generate matrix covariates $X$ and vector covariates $z$, both of which contain pseudorandom values drawn from the standard normal distribution. Then, we generate the true matrix coefficient $B = B_1 B_2^T$, where $B_1 \in \mathbb{R}^{m \times r}$, $B_2 \in \mathbb{R}^{q \times r}$, which $B_1$ and $B_2$ also satisfy standard normal distributions, and $r$ controls the rank of $B$. Thus, we obtain the low-$r$-rank matrix $B$. In addition, we randomly generate the true coefficient $\gamma$ where its elements are random numbers chosen from a binomial distribution with parameters 1 and 0.5. Moreover, the percentage of non-zeros entries is controlled by a sparsity level $s$. Finally, we can get the observation data $y$ by $y = \langle X, B \rangle + \langle z, \gamma \rangle + \varepsilon$ with $\varepsilon \sim n(0, \sigma^2 I)$. We test the case $r = 5, 10, 15, 20, 30$ and $s = 0.01, 0.05, 0.1, 0.2, 0.5$ for $(m, n, p, q) = (100, 50, 20000, 30)$.

Table 1 presents the comparison results for $r = 5, 10, 15, 20, 30$, respectively. In the tables, RMSE denotes the root mean squared errors. Moreover, the best method for each test case is highlighted in boldface. Some observations can be made: (1) No matter for $r = 5, 10, 15, 20, 30$ and $s = 0.01, 0.05, 0.1, 0.2, 0.5$, the performance of NonMM is always the best in terms of RMSE($B$) and RMSE($\gamma$), which illustrates that our proposed method is the most effective. For example, for $r = 5$, $s = 0.5$, the gain of RMSE($B$) is 0.1139 when compared with LMM, and 0.1731 when compared with LSMM. The gain of RMSE($\gamma$) is 0.0279 when compared with LMM, and 0.9776 when compared with LSMM. (2) Compared with LMM, LSMM and NonMM reduce RMSE($\gamma$) dramatically. This is because that LSMM and NonMM apply the sparse regularization. Moreover, compared with LSMM, NonMM always achieves the smaller RMSE($\gamma$). The superior is due to the nonconvex regularization, which exploits the structure property.

4.2. **Comparison results for Real-world data.** In this subsection, we test the trip time prediction from partial trajectories. Estimation of travel time for taxis is of great importance for existing electronic dispatch systems. It is an accurate dataset describing the trajectories performed by all the 442 taxis running in the city of Porto for a period of one year. We choose partial dataset with 7733 trajectories,
which every trajectory includes multiple features. Record the latitude and longitude coordinates every 15 seconds while the taxi is running. Since the longest driving time measures the position of 922 times, the short travel time fills the corresponding ending position according to the longest distance. It also contains 7 regular variates, such as, trip id, call type, origin call, origin stand, taxi id, day type and start timestamp. Trip id contains a unique identifier for 15 trips. Call type identifies the way used to demand this service, which contains three types, i.e., this trip was dispatched from the central, demanded directly to a taxi driver at a specific stand and otherwise (demanded on a random street). Origin call contains a unique identifier for each phone number which was used to demand. Origin stand contains a unique identifier for the taxi stand. Taxi id contains a unique identifier for 404 taxi drivers that performed each trip. Day type identifies the day type of the trip’s start, i.e., holiday or any other special day, a day before holiday and otherwise. Start timestamp identifies the trip’s start.

Thus, $X$ is a $922 \times 2$ matrix and $z$ is a 7 regular vector. Furthermore, we removed 32 trajectories due to the missing observation of coordinates. Our aim is to extract

| Cases | RMSE(B) | RMSE(γ) |
|-------|---------|---------|
|       | LMM     | LSMM    | NonMM  | LMM     | LSMM    | NonMM  |
| r      | s       |         |        |         |         |        |
| 0.01   | 0.0000  | 0.0000  | 0.0000 | 0.0000  | 0.1806  | 0.1421  | 0.1101 |
| 0.05   | 0.0456  | 0.0437  | 0.0274 | 0.0437  | 0.6453  | 0.2742  | 0.2336 |
| 5      | 0.1     | 0.0905  | 0.0688 | 0.0496  | 0.9162  | 0.3788  | 0.3176 |
| 0.2    | 0.1314  | 0.1082  | 0.0911 | 0.0911  | 1.0822  | 0.5063  | 0.4410 |
| 0.5    | 0.2744  | 0.2152  | 0.1013 | 0.2152  | 1.5868  | 0.7371  | 0.7092 |
| 0.01   | 0.0000  | 0.0000  | 0.0000 | 0.0000  | 0.1927  | 0.1513  | 0.1217 |
| 0.05   | 0.0781  | 0.0511  | 0.0382 | 0.0382  | 0.7663  | 0.3312  | 0.2961 |
| 10     | 0.1     | 0.1072  | 0.0749 | 0.0536  | 1.0452  | 0.3879  | 0.3401 |
| 0.2    | 0.1339  | 0.1217  | 0.1161 | 0.1217  | 1.2696  | 0.5305  | 0.4918 |
| 0.5    | 0.2947  | 0.2440  | 0.1278 | 0.2440  | 1.9561  | 0.7971  | 0.7330 |
| 0.01   | 0.0000  | 0.0000  | 0.0000 | 0.0000  | 0.2197  | 0.1920  | 0.1405 |
| 0.05   | 0.0982  | 0.0749  | 0.0654 | 0.0654  | 0.8408  | 0.3223  | 0.2893 |
| 15     | 0.1     | 0.1323  | 0.1130 | 0.1108  | 1.2501  | 0.3534  | 0.3293 |
| 0.2    | 0.1846  | 0.1560  | 0.1272 | 0.1272  | 1.6730  | 0.4988  | 0.4494 |
| 0.5    | 0.3176  | 0.2812  | 0.2176 | 0.2176  | 2.3608  | 0.7817  | 0.7502 |
| 0.01   | 0.0000  | 0.0000  | 0.0000 | 0.0000  | 0.1936  | 0.1620  | 0.1509 |
| 0.05   | 0.0987  | 0.0791  | 0.0732 | 0.0732  | 0.9081  | 0.3168  | 0.2719 |
| 20     | 0.1     | 0.1528  | 0.1454 | 0.1402  | 1.416   | 0.4583  | 0.3961 |
| 0.2    | 0.1904  | 0.1821  | 0.1723 | 0.1723  | 1.8304  | 0.4722  | 0.4148 |
| 0.5    | 0.3876  | 0.2922  | 0.2134 | 0.2134  | 2.4961  | 0.8049  | 0.7891 |
| 0.01   | 0.0000  | 0.0000  | 0.0000 | 0.0000  | 0.1768  | 0.1620  | 0.1400 |
| 0.05   | 0.1071  | 0.0932  | 0.0858 | 0.0858  | 1.0089  | 0.2360  | 0.2232 |
| 30     | 0.1     | 0.1608  | 0.1572 | 0.1469  | 1.5286  | 0.3241  | 0.3062 |
| 0.2    | 0.2487  | 0.2371  | 0.2206 | 0.2206  | 2.082   | 0.4920  | 0.4599 |
| 0.5    | 0.4170  | 0.3943  | 0.3725 | 0.3725  | 2.6203  | 0.7856  | 0.8041 |
features to predict taxi’s travel time for complete journeys. We summarize the comparison Results of RMSE(y) in Table 2.

| Rate  | LMM  | LSMM | NonMM |
|-------|------|------|-------|
| 5-fold | 0.1321 | 0.1178 | **0.1025** |
| 10-fold | 0.1972 | 0.1630 | **0.1589** |

Here, we divide data into a training and a testing sample by using 5-fold or 10-fold Cross-Validation. It is shown that NonMM enjoys a smaller RMSE(y) than LMM and LSMM.

5. **Conclusion.** In this paper, we propose a novel nonconvex mixed matrix minimization model. The key feature of our model is the nonconvex regularizations, which can achieve much more accurate than the convex low-rank and sparse regularizations. We also demonstrate that our proposed nonconvex mixed matrix minimization could be easily solved by the nonconvex ALM with convergence analysis. The accuracy and efficiency of the nonconvex mixed matrix minimization model is finally verified by numerical examples. Naturally, many questions need a deeper exploration. For future research, an interesting direction is to integrate nonconvex functions with nuclear norm, and the improvement may be better.

**Acknowledgments.** The authors would like to thank the anonymous referees for their comments and suggestions, which have greatly improved the paper.

**REFERENCES**

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via alternating direction method of multipliers, *Foundations and Trends® in Machine Learning*, 3 (2011), 1–122.
[2] J. Cai, E. Candès and Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization*, 20 (2010), 1956–1982.
[3] X. Chen, M. Ng and C. Zhang, Non-lipschitz $\ell_p$-regularization and box constrained model for image restoration, *IEEE Transactions on Image Processing*, 21 (2012), 4709–4721.
[4] Y. Chen, N. Xiu and D. Peng, Global solutions of non-Lipschitz $S_2 - S_p$ minimization over the positive semidefinite cone, *Optimization Letters*, 8 (2014), 2053–2064.
[5] D. Donoho, De-noising by soft-thresholding, *IEEE Transactions on Information Theory*, 41 (1995), 613–627.
[6] M. Elad, Why simple shrinkage is still relevant for redundant representations?, *IEEE Transactions on Information Theory*, 52 (2006), 5559–5569.
[7] J. Fan, Comments on “wavelets in statistics: A review” by A. Antoniadis, *Journal of the Italian Statistical Society*, 6 (1997), 191–138.
[8] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96 (2001), 1348–1360.
[9] J. Fan, L. Xue and H. Zou, Strong oracle optimality of folded concave penalized estimation, *Annals of Statistics*, 42 (2014), 819–849.
[10] M. Fazel, T. Pong, D. Sun and P. Tseng, Hankel matrix rank minimization with applications to system identification and realization, *SIAM Journal on Matrix Analysis and Applications*, 34 (2013), 946–977.
[11] D. Gabay, Chapter ix applications of the method of multipliers to variational inequalities, *Studies in Mathematics and Its Applications*, 15 (1983), 299–331.
[12] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Computers and Mathematics with Applications*, 2 (1976), 17–40.
[13] M. Hong, Z. Luo and M. Razaviyayn, Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems, *SIAM Journal on Optimization*, **26** (2016), 337–364.

[14] G. Li and T. Peng, Douglass-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems, *Mathematical Programming*, **159** (2016), 371–401.

[15] S. Negahban and M. Wainwright, Estimation of (near) low-rank matrices with noise and high-dimensional scaling, *Annals of Statistics*, **39** (2011), 1069–1097.

[16] M. Nikolova, M. Ng, S. Zhang and W. Ching, Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization, *SIAM Journal on Imaging Sciences*, **1** (2008), 2–25.

[17] G. Obozinski, M. Wainwright and M. Jordan, Support union recovery in high-dimensional multivariate regression, *Annals of Statistics*, **39** (2011), 1–47.

[18] L. Rudin and S. Osher, Total variation based image restoration with free local constraints, *In Proceedings of the IEEE International Conference on Image Processing*, **1** (1994), 31–35.

[19] R. Rockafellar and R. Wets, *Variational Analysis*, Springer Science and Business Media, 2009.

[20] P. Shang and L. Kong, On the degrees of freedom of mixed matrix regression, *Mathematical Problems in Engineering*, **2017** (2017), Art. ID 694265, 8 pp.

[21] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, **58** (1996), 267–288.

[22] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight, Sparsity and smoothness via the fused Lasso, *Journal of the Royal Statistical Society, Series B*, **67** (2005), 91–108.

[23] F. Wang, W. Cao and Z. Xu, Convergence of multi-block Bregman ADMM for nonconvex composite problems, *Science China Information Sciences*, **61** (2018), 122101, 12pp.

[24] X. Xiu, L. Kong, Y. Li and H. Qi, Iterative reweighted methods for $\ell_1 - \ell_p$ minimization, *Computational Optimization and Applications*, **70** (2018), 201–219.

[25] X. Xiu, W. Liu, L. Li and L. Kong, Alternating direction method of multipliers for nonconvex fused regression problems, *Computational Statistics and Data Analysis*, **136** (2019), 59–71.

[26] L. Yang, T. Pong and X. Chen, Alternating direction method of multipliers for a class of nonconvex and nonsmooth problems with applications to background/foreground extraction, *SIAM Journal on Imaging Sciences*, **10** (2017), 74–110.

[27] M. Yuan, A. Ekici, Z. Lu and R. Monteiro, Dimension reduction and coefficient estimation in multivariate linear regression, *Journal of Royal Statistical Society, Series B*, **69** (2007), 329–346.

[28] C. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, **38** (2010), 894–942.

[29] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, *Journal of Machine Learning Research*, **11** (2010), 1081–1107.

[30] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, **67** (2005), 301–320.

[31] H. Zhou and L. Li, Regularized matrix regression, *Journal of the Royal Statistical Society, Series B*, **76** (2014), 463–483.

Received for publication February 2019.

E-mail address: meijuanshang@163.com
E-mail address: 16121656@bjtu.edu.cn
E-mail address: lchkong@bjtu.edu.cn
E-mail address: xxxiu@bjtu.edu.cn
E-mail address: yy@pku.edu.cn