Electronic Journal of Statistics
Vol. 15 (2021) 1909–1950
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1829

Double fused Lasso regularized regression with both matrix and vector valued predictors[*]

Mei Li[†] **and Lingchen Kong**

Department of Applied Mathematics, Beijing Jiaotong University
e-mail: 18118016@bjtu.edu.cn; lchkong@bjtu.edu.cn

Zhihua Su

Department of Statistics, University of Florida
e-mail: zhihuasu@ufl.edu

Abstract: In many contemporary applications such as longitudinal studies, neuroimaging or civil engineering, a dataset can contain high dimensional measurements on both matrix-valued and vector-valued variables. Such structure demands statistical tools that can extract information from both types of measurements. In this paper, we propose a double fused Lasso regularized method to handle both matrix-valued and vector-valued predictors under the context of linear regression and logistic regression. An efficient and scalable sGS-ADMM (symmetric Gauss-Seidel based alternating direction method of multipliers) algorithm is derived to obtain the estimator. Global convergence and the Q-linear rate of convergence for the algorithm is established. Consistency of the double fused Lasso estimators holds under mild conditions. Numerical experiments and examples show that the double fused Lasso estimators achieve efficient gains in estimation and better prediction performance compared to existing estimators.

MSC2020 subject classifications: Primary 62J05, 62J12, 62F12; secondary 90C25.
Keywords and phrases: Lasso, matrix-variate regression, Q-linear rate, risk bound, sGS-ADMM.

Received February 2020.

Contents

[*] This work was supported in part by the National Natural Science Foundation of China (12071022), grant (632688) from Simons Foundation and the 111 Project of China (B16002).

[†] Corresponding author

1. Introduction

In the era of big data, many datasets with complex structures emerge, which may contain matrix-valued and vector-valued variables simultaneously. For example, bike sharing schemes have gained increasing popularity in the recent years and become an integrated part of transportation network in many cities. Bike rental demand depends on the weather conditions and social factors [13], and a method to estimate or forecast the demand is important for bike sharing systems. The two-year historical log from Capital Bike sharing system in Washington D.C. contains daily bike rental counts from January 1st, 2011 to December 31th, 2012. It also includes a 24×6 matrix containing hourly weather information such as temperature, humidity and wind speed for each day. Additional information such as month, year, days of the week and holiday indicator is also recorded, which constitutes a vector-valued predictor. The daily bike rental count is taken to be the response. Multiple linear regression has been used to tackle such data [15] with vector-valued predictors. Since we have both matrix-valued predictors (weather conditions) and vector-valued predictors, traditional regression methods are not directly applicable. New regression tools that can be adapted to such data structure are in need. [50] proposed a matrix regression model

$$y = \langle X, B \rangle + \langle z, \gamma \rangle + \varepsilon, \tag{1.1}$$

where $y \in \mathbb{R}$ is a continuous response, $X \in \mathbb{R}^{m\times q}$ is a matrix-valued predictor and $z \in \mathbb{R}^p$ is a vector-valued predictor. The matrix $B \in \mathbb{R}^{m\times q}$ is a coefficient matrix with the same size as X and $\gamma \in \mathbb{R}^p$ contains the coefficients for z. The inner product $\langle X, B\rangle$ is defined as $\mathrm{tr}(X^T B)$. The $\varepsilon \in \mathbb{R}$ is the noise. Without the matrix-valued predictor X, (1.1) reduces to the standard linear regression $y = \langle z, \gamma\rangle + \varepsilon$. Without the vector-valued predictor z, (1.1) reduces to the matrix regression model with only a matrix-valued predictor $y = \langle X, B\rangle + \varepsilon$.

In some applications, response variable can be binary. For example, the diabetes dataset contains physical exam information of 2476 staffs of Beijing Jiaotong University from 2016 to 2018. During each annual physical exam, 62 features are measured including concentration and volume of erythrocytes, leukocytes and platelets, blood sugar concentration, kidneys and liver function tests, facial features and dietary preferences, giving a 62 by 3 matrix of physical exam results. In addition, seven covariates including gender, education, occupation, disability status are recorded for each patient. In 2018, 237 staffs are diagnosed to have diabetes, and 2239 staffs do not have diabetes. The association between diabetes and potential predictors is of special interest. In this case, we have a logistic regression model with both matrix-valued predictor (physical exam results) and vector-valued predictor. The response is the diabetic indicator, which takes value 1 if the patient has diabetes and 0 otherwise. The logistic regression model is formulated as

$$\mathrm{logit} P(y=1) = \log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \langle X, B\rangle + \langle z, \gamma\rangle, \tag{1.2}$$

where $X \in \mathbb{R}^{m\times q}$ is a matrix-valued predictor and $z \in \mathbb{R}^p$ is a vector-valued predictor. The matrix $B \in \mathbb{R}^{m\times q}$ contains the coefficients for the matrix-valued predictor X and $\gamma \in \mathbb{R}^p$ contains the coefficients for vector-valued predictor z. The regular logistic regression and matrix variate logistic regression are both special cases of (1.2) without the X term or z term respectively.

Since the complex structure of models (1.1) and (1.2) and high-dimensionality in many applications, it is common to assume that the coefficients in (1.1) and (1.2) have sparse structure, or can be approximated by low-rank structure. To induce such structure, a variety of regularization methods have recently been proposed. Under the context of linear regression (1.1), if the regression model only has vector-valued predictors, popular penalization methods include the Lasso [44], the fused Lasso [45], the elastic net [53], and smooth clipped absolute deviation (SCAD) [12] are proposed. Regularization methods for matrix-valued parameters include, but not limit to, the nuclear norm regularization [8, 30, 34], multivariate group Lasso [36], multivariate sparse group Lasso [28] and matrix regression model based on singular values [50]. Under the logistic regression context (1.2), if we only have the vector-valued predictor, [1], [38] and [43] introduced the logistic regression model with sparse constraints. [31] dealt with the group Lasso for logistic regression model. [22] proposed the matrix-variate logistic regression for data only containing a matrix-valued variable.

In this paper, we propose a double fused Lasso regularized method which imposes a nuclear norm and a fused Lasso norm on the rows of B. This induces a low rank structure in B as well as the sparsity in the difference of successive rows, since in a longitudinal or imaging processing application, coefficients may change smoothly over a particular period of time. An L_1 norm and a fused Lasso norm is imposed on γ, which encourages sparsity in γ as well as the difference of successive elements. The formulation is then

$$\begin{aligned}\min_{B,\gamma} \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle X_i, B\rangle - \langle z_i, \gamma\rangle)^2 + \lambda_1\|B\|_* + \lambda_2 \sum_{j=2}^{m}\|B_{j\cdot} - B_{(j-1)\cdot}\|_1 \\ +\lambda_3\|\gamma\|_1 + \lambda_4\sum_{k=2}^{p}|\gamma_k - \gamma_{k-1}|,\end{aligned} \tag{1.3}$$

under the linear regression model (1.1). Under the logistic regression model (1.2), it is

$$\begin{aligned}\min_{B,\gamma} \sum_{i=1}^{n}\log(1 + e^{\langle X_i,B\rangle + \langle z_i,\gamma\rangle}) - y_i(\langle X_i, B\rangle + \langle z_i, \gamma\rangle) + \lambda_1\|B\|_* \\ +\lambda_2 \sum_{j=2}^{m}\|B_{j\cdot} - B_{(j-1)\cdot}\|_1 + \lambda_3\|\gamma\|_1 + \lambda_4\sum_{k=2}^{p}|\gamma_k - \gamma_{k-1}|.\end{aligned} \tag{1.4}$$

Here γ_k denotes the kth element in γ, $B_{j\cdot}$ denotes the jth row in B and n denotes the sample size. We impose the L_1 norm $\|\gamma\|_1$ and fused Lasso term $\sum_{k=2}^{p}|\gamma_k - \gamma_{k-1}|$ on the coefficients γ of the vector-valued predictors, and impose the nuclear norm $\|B\|_*$ and matrix-type fused Lasso term $\sum_{j=2}^{m}\|B_{j\cdot} - B_{(j-1)\cdot}\|_1$ on the coefficients B of the matrix-valued predictors. We call model (1.3) as double fused Lasso regularized matrix regression, or DFMR for simplicity. DFMR degenerates to the fused Lasso [45] when $B = 0$. It becomes the matrix-type fused Lasso with $\gamma = 0$, which is an extension of the regularized matrix regression [8, 10, 50]. We call model (1.4) as double fused Lasso regularized matrix logistic regression, or DFMLR for simplicity. DFMLR degenerates to the fused Lasso regularized logistic regression [29] when $B = 0$. It becomes the matrix-type fused Lasso regularized logistic regression with $\gamma = 0$.

To solve similar optimization problem as in (1.3) or (1.4), first-order methods such as alternating direction method of multipliers (ADMM) and augmented Lagrangian method (ALM) are widely used, see e.g., [17, 21]. Specifically, [26] proposed linearized ADMM algorithm for sparse group Lasso and fused Lasso model. [48] considered ALM as a solver for the fused Lasso signal approximator problem. [49] developed an efficient ALM for large-scale non-overlapping sparse group Lasso problems. [50] proposed the Nesterov optimal gradient method for spectral regularized matrix regression. Due to coupled variables in the double fused Lasso penalties, a natural choice for solving the (1.3) and (1.4) is the ADMM algorithm. It is more efficient than ALM because ADMM solves B or γ alternatively instead of solving B and γ simultaneously. However, in high dimensional scenarios, the inversion of high dimensional matrices is the computation bottleneck for scalability. To reduce the computational cost, we consider the dual of (1.3) and (1.4), where the inversion of high dimensional matrices are avoided in such scenarios.

We propose an efficient and scalable symmetric Gauss-Seidel based ADMM (sGS-ADMM) algorithm to solve the dual of (1.3) and (1.4). In particular, every subproblem for dual of (1.3) has a closed-form solution. The global convergence and Q-linear rate of convergence of the algorithm is established. The resulting DFMR or DFMLR estimators show a superior estimation and prediction performance compared with the matrix Lasso and Lasso methods [50]. We also investigate the theoretical properties of the DFMR and DFMLR estimators.

The rest of the paper is organized as follows. Section 2 proposes an efficient sGS-ADMM algorithm to obtain the DFMR and DFMLR estimators, and establishes the global convergence and Q-linear rate of convergence of the algorithm. In Section 3, we investigate the consistency for the DFMR and DFMLR estimators. Section 4 demonstrates the performance of the DFMR estimator and the DFMLR estimator through simulations. Examples are included in Section 5. We conclude this paper and discuss future directions in Section 6. Proofs of theorems and other technical details are deferred to the Appendix.

We first introduce some notations which are useful for further discussion. Given a vector $x \in \mathbb{R}^n$, its L_1 norm, L_2 norm and L_∞ norm are defined as $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\| = \|x\|_2 = \sqrt{\langle x, x\rangle}$, $\|x\|_\infty = \max\{|x_i|, i = 1, 2, \cdots, n\}$. The closed balls centered at 0 with radius $r \geq 0$ based on L_2 norm and L_∞ norm are defined by $B_{\|\cdot\|_2(0;r)} = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq r\}$ and $B_{\|\cdot\|_\infty(0;r)} = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq r\}$. Given a nonempty closed convex set $\Omega \subset \mathbb{R}^n$, define its indicator function by $\delta(x;\Omega) = 0$ if $x \in \Omega$, $\delta(x;\Omega) = \infty$ if $x \notin \Omega$. The distance function from x to Ω is $d(x;\Omega) = \inf\{\|x - \omega\| \mid \omega \in \Omega\}$. The Euclidean projection of x onto Ω is $\Pi(x;\Omega) = \{\omega \in \Omega \mid \|x - \omega\| = d(x;\Omega)\}$. For random variable x, we denote its sub-Gaussian norm as $\|x\|_{\psi_2} = \sup_{p\geq 1}(\mathbb{E}|x|^p)^{1/p}/\sqrt{p}$.

For any matrix $A \in \mathbb{R}^{m\times n}$, its singular value decomposition is denoted by $A = U\Sigma V^{\mathrm{T}}$, where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ have orthonormal columns, r is the rank of A, and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix. The diagonal elements of Σ are called the singular values of A, and we denote them by $\sigma_i(A), i = 1, 2, \cdots, r$. The sub-differential of $\|A\|_*$ is $\partial\|A\|_* = \{UV^{\mathrm{T}} + W | W \in \mathbb{R}^{m\times n}, U^{\mathrm{T}}W = 0, WV = 0, \|W\|_2 \leq 1\}$. The Frobenius norm ($F$-norm), nuclear norm and spectral norm of A are defined as $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$, $\|A\|_* = \sum_{i=1}^r \sigma_i(A)$, $\|A\|_2 = \max\{\sigma_i(A), i = 1, 2, \cdots, r\}$. We use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and smallest eigenvalues of A, respectively. The closed ball centered at 0 with radius $r \geq 0$ based on spectral norm is defined as $B_{\|\cdot\|_2(0;r)} = \{A \in \mathbb{R}^{m\times n} \mid \|A\|_2 \leq r\}$. For symmetric matrices A_1 and A_2, $A_1 \succeq A_2$ means that $A_1 - A_2$ is positive semidefinite, and $A_1 \succ A_2$ means that $A_1 - A_2$ is positive definite.

2. Estimation algorithm

In this section, we propose an efficient sGS-ADMM algorithm to obtain the DFMR estimator, and then generalize the algorithm to solve for the DFMLR estimator. Convergence of the algorithm is discussed.

2.1. The sGS-ADMM algorithm for DFMR

2.1.1. Model reformulation and dual

We first reformulate the model (1.3) to find its dual. Let (y_i, X_i, z_i) be independent and identical copies of (y, X, z), $i = 1, \dots, n$. For simplicity, we denote $y = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}, \mathbb{X} = (\mathrm{vec}(X_1), \cdots, \mathrm{vec}(X_n))^{\mathrm{T}}$ and $\mathbb{Z} = (z_1, \cdots, z_n)^{\mathrm{T}}$, where vec operator stacks a matrix into a vector columnwise. Let A_i be an $(i-1) \times i$ matrix that has the following structure

$$A_i = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \tag{2.1}$$

We define matrix $C \in \mathbb{R}^{(m-1)q \times mq}$ as

$$C = \begin{pmatrix} A_m & 0 & \cdots & 0 \\ 0 & A_m & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{pmatrix}. \tag{2.2}$$

Then (1.3) can be reformulated as

$$\begin{aligned} \min_{B,\gamma} \frac{1}{2}\|y - \mathbb{X}\mathrm{vec}(B) - \mathbb{Z}\gamma\|_2^2 + \lambda_1\|B\|_* + \lambda_2\|C\mathrm{vec}(B)\|_1 \\ + \lambda_3\|\gamma\|_1 + \lambda_4\|A_p\gamma\|_1. \end{aligned} \tag{2.3}$$

We introduce two slack variables $\xi \in \mathbb{R}^n, \eta \in \mathbb{R}^{(m-1)q}$ with $\xi = y - \mathbb{X}\mathrm{vec}(B) - \mathbb{Z}\gamma$ and $\eta = C\mathrm{vec}(B)$. Then (2.3) can be written as

$$\begin{aligned} \min_{B,\gamma,\xi,\eta} \ & \frac{1}{2}\|\xi\|_2^2 + \lambda_1\|B\|_* + \lambda_2\|\eta\|_1 + \lambda_3\|\gamma\|_1 + \lambda_4\|A_p\gamma\|_1 \\ s.t. \quad & y - \mathbb{Z}\gamma - \mathbb{X}\mathrm{vec}(B) = \xi, \\ & C\mathrm{vec}(B) = \eta. \end{aligned} \tag{2.4}$$

The objective function in (2.4) is convex with respect to every variable B, γ, ξ and η, but it is a smooth function only with respect to ξ. The two constraints are both linear. So (2.4) is a convex and nonsmooth optimization problem. Let $P(\gamma) = \lambda_3\|\gamma\|_1 + \lambda_4\|A_p\gamma\|_1$. The Lagrangian function of (2.4) is

$$\begin{aligned} \mathcal{L}(B, \gamma, \xi, \eta; u, v) = & \frac{1}{2}\|\xi\|_2^2 + \lambda_1\|B\|_* + \lambda_2\|\eta\|_1 + P(\gamma) \\ & - \langle u, \xi - (y - \mathbb{X}\mathrm{vec}(B) - \mathbb{Z}\gamma)\rangle - \langle v, C\mathrm{vec}(B) - \eta\rangle, \end{aligned}$$

where $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^{(m-1)q}$ are Lagrange multipliers. Let $P^*(\cdot)$ be the Fenchel conjugate function of $P(\cdot)$. Then the dual of (2.4) is equivalent to

$$\min_{u,v,D,w,t} \frac{1}{2}\|u\|_2^2 - \langle u, y\rangle + P^*(w) + \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) + \delta(D; B_{\|\cdot\|_2(0;\lambda_1)})$$

$$
\begin{aligned}
s.t.\quad & \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D) = 0,\\
& \mathbb{Z}^{\mathrm{T}}u \qquad\quad - w \qquad = 0,\\
& \qquad\qquad v + t \qquad\quad = 0.
\end{aligned}
\tag{2.5}
$$

Based on the duality theorem, we can use the ADMM algorithm to get the solutions of (2.4) and (2.5). The augmented Lagrangian function is given by

$$
\begin{aligned}
\mathcal{L}_\sigma(B,\gamma,\xi,\eta;u,v) =& \frac{1}{2}\|\xi\|_2^2 + \lambda_1\|B\|_* + \lambda_2\|\eta\|_1 + P(\gamma)\\
&- \langle u, \xi - (y - \mathbb{X}\mathrm{vec}(B) - \mathbb{Z}\gamma)\rangle - \langle v, C\mathrm{vec}(B) - \eta\rangle\\
&+ \frac{1}{2}\|\xi - (y - \mathbb{X}\mathrm{vec}(B) - \mathbb{Z}\gamma)\|^2 + \frac{1}{2}\|C\mathrm{vec}(B) - \eta\|^2.
\end{aligned}
$$

If we directly optimize $\mathcal{L}_\sigma(B,\gamma,\xi,\eta;u,v)$ using the ADMM algorithm, it involves the computation of the inverses of $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ and $\mathbb{Z}^{\mathrm{T}}\mathbb{Z}$, which can be computationally expensive in high dimensional problems. Furthermore, note that (2.4) is a multi-block convex composite optimization problem with linear equality constraints. For such problems, convergence cannot be achieved by ADMM in general [4]. An augmented ADMM algorithm is proposed in [52]. However, its convergence is guaranteed only for two-block optimization problems and it may not achieve convergence for multi-block optimization problems. Therefore we consider the optimization problem (2.5), and employ the sGS-ADMM algorithm to solve (2.5). The sGS-ADMM algorithm is first proposed in [5], which combines the sGS technique with ADMM algorithm. A brief introduction on the sGS technique and sGS-ADMM algorithm for a general convex composite programming model is included in Appendix A.2.

2.1.2. Algorithm analysis

Now we employ the sGS-ADMM algorithm to solve the optimization problem (2.5). For convenience, let $\alpha = (\mathrm{vec}(D)^{\mathrm{T}}, w^{\mathrm{T}}, t^{\mathrm{T}})^{\mathrm{T}}$. Although (2.5) contains five variables u, v, D, w, t, it can be considered as a 3-block optimization problem with u, v and α, and only contains a nonsmooth term involving the variable α. The augmented Lagrangian function is

$$
\begin{aligned}
\mathcal{L}_\sigma(u,v,\alpha;x) =& \frac{1}{2}\|u\|_2^2 - \langle u, y\rangle + P^*(w) + \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) + \delta(D; B_{\|\cdot\|_2(0;\lambda_1)})\\
&- \langle x_1, \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D)\rangle - \langle x_2, \mathbb{Z}^{\mathrm{T}}u - w\rangle - \langle x_3, v + t\rangle\\
&+ \frac{\sigma}{2}\|\mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D)\|^2 + \frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}}u - w\|^2 + \frac{\sigma}{2}\|v + t\|^2,
\end{aligned}
$$

where $\sigma > 0$, $x_1 \in \mathbb{R}^{mq}$, $x_2 \in \mathbb{R}^p$ and $x_3 \in \mathbb{R}^{(m-1)q}$ are Lagrange multipliers, and $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}}, x_3^{\mathrm{T}})^{\mathrm{T}}$.

Note that the augmented Lagrangian function is strongly convex with respect to every variable u, v and α. Hence, the majorization step given in [5] is not necessary. Moreover, every subproblem in the sGS-ADMM algorithm has a

Table 1

Iterative scheme of the sGS-ADMM algorithm for solving (2.5)

```
Algorithm 1:
    Input: $X, Z, y$ and tolerance level $tol$. Choose $\lambda_1 > 0, \lambda_2 > 0, \lambda_3 > 0, \lambda_4 > 0$ and $\sigma > 0$.
    Let $\tau \in (0, (1+\sqrt{5})/2)$ be the step-length. Set the initial point $(u^0, v^0, \alpha^0, x^0)$.
    For $k = 0, 1, \cdots$, perform the following steps:
    Step 1a. (Backward GS sweep) Compute $u^{k+\frac{1}{2}}$ and $v^{k+\frac{1}{2}}$,
      $u^{k+\frac{1}{2}} = \arg\min_u \mathcal{L}_\sigma(u, v^k, \alpha^k; x^k)$,
      $v^{k+\frac{1}{2}} = \arg\min_v \mathcal{L}_\sigma(u^{k+\frac{1}{2}}, v, \alpha^k; x^k)$.
    Step 1b. (Forward GS sweep) Compute $u^{k+1}$ , $v^{k+1}$ and $\alpha^{k+1}$,
      $\alpha^{k+1} = \arg\min_\alpha \mathcal{L}_\sigma(u^{k+\frac{1}{2}}, v^{k+\frac{1}{2}}, \alpha; x^k)$,
      $v^{k+1} = \arg\min_v \mathcal{L}_\sigma(u^{k+\frac{1}{2}}, v, \alpha^{k+1}; x^k)$,
      $u^{k+1} = \arg\min_u \mathcal{L}_\sigma(u, v^{k+1}, \alpha^{k+1}; x^k)$.
    Step 2. Update Lagrange multipliers $x_1^{k+1}, x_2^{k+1}$ and $x_3^{k+1}$,
      $x_1^{k+1} = x_1^k - \tau\sigma(\mathbb{X}^{\mathrm{T}} u^{k+1} + C^{\mathrm{T}} v^{k+1} - \mathrm{vec}(D^{k+1}))$,
      $x_2^{k+1} = x_2^k - \tau\sigma(\mathbb{Z}^{\mathrm{T}} u^{k+1} - w^{k+1})$,
      $x_3^{k+1} = x_3^k - \tau\sigma(v^{k+1} + t^{k+1})$.
      If $eta < tol = 10^{-3}$ stop
```

closed-form solution. Then given the values of u, v and α after the kth iteration u^k, v^k, and α^k, the iterative scheme of the sGS-ADMM algorithm for solving (2.5) is summarized in Table 1.

Note that each subproblem in Table 1 has a closed-form solution, which is due to the properties of the augmented Lagrangian function or the properties of proximal mapping (see Appendix A.1 for properties of proximal mapping and Appendix A.3 for derivation of the closed-form solution of each subproblem). The updates of the Lagrange multipliers x_1^{k+1}, x_2^{k+1} and x_3^{k+1} are straightforward as given in Step 2 in Table 1. The stopping criterion *eta* is derived from the KKT condition. See Appendix A.3 for details.

2.1.3. Convergence analysis

In this section, we establish the global convergence and Q-linear rate of convergence of the sGS-ADMM algorithm in Table 1. While general results for the convergence of sGS-ADMM algorithm are available in [5] and [20], we verify that the assumptions for the global convergence and Q-linear rate of convergence are satisfied in our context, see Appendix A.5 for details. Let $\theta^k = (u^k, v^k, \alpha^k, x^k)$ be the value of θ after the kth iteration.

Theorem 2.1. *The sequence $\{u^k, v^k, \alpha^k, x^k\}$ converges to the optimal solution $(\bar{u}, \bar{v}, \bar{\alpha}, \bar{x})$. Moreover, $(\bar{u}, \bar{v}, \bar{\alpha})$ is an optimal solution of (2.5), and $\bar{x}$ is an optimal solution of (2.4).*

To investigate the convergence rate, we first give a brief review on Q-linear rate of convergence. Let $\{x^k\}$ be the sequence of iterates and x^* the optimal solution. Suppose that

$$\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^r} = q, \qquad r \geq 1,$$

we say that the Q-order of convergence of $\{x^k\}$ to x^* is r. In particular, if $r = 1$ and $0 < q < 1$, then the convergence is said to be Q-linear. If $r = 1$ and $q = 0$, the convergence is Q-superlinear. The prefix "Q" means "quotient". Another type of linear convergence is R-linear convergence, where the prefix "R" is for "root". We say that $\{x^k\}$ converges to x^* R-linearly if

$$\|x^k - x^*\| \leq \nu_k$$

for all k, and ν_k converges Q-linearly to zero. More details are included in [35].

Now we introduce some notations. For any self-adjoint positive semi-definite linear operator $\mathcal{M}_0 : \mathcal{X} \to \mathcal{X}$, let $dist_{\mathcal{M}_0}(x, \Omega) = \inf_{x'\in\Omega} \|x' - x\|_{\mathcal{M}_0}$ for any $x \in \mathcal{X}$ and any set $\Omega \in \mathcal{X}$. Recall that $\alpha = (\text{vec}(D)^T, w^T, t^T)^T$, we define $h(\alpha) = P^*(w) + \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) + \delta(D; B_{\|\cdot\|_2(0;\lambda_1)})$. According to Theorem 12.17 in [41], there exists a self-adjoint and positive semi-definite linear operator Σ_α such that for all α, α', $\zeta \in \partial h(\alpha)$ and $\zeta' \in \partial h(\alpha')$, $\langle \zeta' - \zeta, \alpha' - \alpha\rangle \geq \|\alpha' - \alpha\|^2_{\Sigma_\alpha}$. Let Φ be a linear operator such that for all (u, v, α, x), its adjoint is

$$\Phi^*(u, v, \alpha, x) = (\mathbb{X}, \mathbb{Z}, 0)^{\mathrm{T}} u + (C, 0, I)^{\mathrm{T}} v - Diag(I_{mq}, I_p, -I_{(m-1)q})\alpha.$$

For any $\tau \in (0, (1+\sqrt{5})/2)$, let $s_\tau = \frac{5-\tau-3\min\{\tau,\tau^{-1}\}}{4}$, and define a self-adjoint linear operator as follows $\mathcal{M} := Diag(I, \sigma(I + CC^{\mathrm{T}}), \sigma\Sigma_I + \Sigma_\alpha, (\tau\sigma)^{-1}I_x) + s_\tau\sigma\Phi\Phi^*$, where $\Sigma_I = I_x = Diag(I_{mq}, I_p, I_{(m-1)q})$. Let $\bar{\Omega}$ be the optimal solution set satisfying the KKT conditions.

Theorem 2.2. *There exists $0 < \mu < 1$ such that for all $k \geq 1$,*

$$dist^2_{\mathcal{M}}(\theta^{k+1}, \bar{\Omega}) \leq \mu dist^2_{\mathcal{M}}(\theta^k, \bar{\Omega}).$$

Theorem 2.2 establishes the Q-linear rate of convergence for the sGS-ADMM algorithm, which guarantees that the sequence $\{u^k, v^k, \alpha^k, x^k\}$ generated by our algorithm converges to the optimal solution.

2.2. *The sGS-ADMM algorithm for DFMLR*

2.2.1. *Model reformulation and dual*

To obtain the DFMLR estimator, we first reformulate the optimization problem (1.4) as

$$\min_{B,\gamma} \sum_{i=1}^n \log(1 + e^{\langle X_i, B\rangle + \langle z_i,\gamma\rangle}) - y_i(\langle X_i, B\rangle + \langle z_i, \gamma\rangle) + \lambda_1\|B\|_* + \lambda_2\|C\text{vec}(B)\|_1 + \lambda_3\|\gamma\|_1 + \lambda_4\|A_p\gamma\|_1, \tag{2.6}$$

where matrices C and A_p are defined in (2.1) and (2.2). Introduce two slack variables $\xi \in \mathbb{R}^n, \eta \in \mathbb{R}^{(m-1)q}$ with $\xi = \mathbb{X}\text{vec}(B) + \mathbb{Z}\gamma$ and $\eta = C\text{vec}(B)$. Then

(2.6) can be written as

$$\begin{aligned}&\min_{B,\gamma,\xi,\eta} \sum_{i=1}^{n} \log(1+e^{\xi_i}) - y_i\xi_i + \lambda_1\|B\|_* + \lambda_2\|\eta\|_1 + P(\gamma)\\ &s.t. \quad \mathbb{X}\mathrm{vec}(B) + \mathbb{Z}\gamma = \xi,\\ &\qquad\ C\mathrm{vec}(B) \qquad = \eta.\end{aligned} \tag{2.7}$$

The objective function in (2.7) is convex and nonsmooth. The dual of (2.7) is equivalent to

$$\begin{aligned}&\min_{u,v,D,w,s,t} \sum_{i=1}^{n}(1-s_i)\log(1-s_i) + s_i\log s_i + P^*(w) + \delta(t;B_{\|\cdot\|_\infty(0;\lambda_2)})\\ &\qquad\qquad +\delta(D;B_{\|\cdot\|_2(0;\lambda_1)})\\ &s.t. \qquad \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D) = 0,\\ &\qquad\quad\ \mathbb{Z}^{\mathrm{T}}u \qquad\quad - w \qquad = 0,\\ &\qquad\quad\ u_i \qquad\quad + s_i \qquad = y_i,\\ &\qquad\qquad\qquad\ v + t \qquad = 0.\end{aligned} \tag{2.8}$$

2.2.2. Algorithm analysis

The optimization problem (2.8) can also be solved by the sGS-ADMM algorithm. It contains six variables u, v, D, w, s, t. Let $\alpha = (\mathrm{vec}(D)^{\mathrm{T}}, w^{\mathrm{T}}, s^{\mathrm{T}}, t^{\mathrm{T}})^{\mathrm{T}}$. The augmented Lagrangian function is

$$\begin{aligned}\mathcal{L}_\sigma(u,v,\alpha;x) =& \sum_{i=1}^{n}(1-s_i)\log(1-s_i) + s_i\log s_i + P^*(w) + \delta(t;B_{\|\cdot\|_\infty(0;\lambda_2)})\\ &+ \delta(D;B_{\|\cdot\|_2(0;\lambda_1)}) - \langle x_1, \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D)\rangle - \langle x_2, \mathbb{Z}^{\mathrm{T}}u - w\rangle\\ &- \langle x_3, u+s-y\rangle - \langle x_4, v+t\rangle + \frac{\sigma}{2}\|\mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D)\|^2\\ &+ \frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}}u - w\|^2 + \frac{\sigma}{2}\|u+s-y\|^2 + \frac{\sigma}{2}\|v+t\|^2,\end{aligned}$$

where $\sigma > 0$, $x_1 \in \mathbb{R}^{mq}$, $x_2 \in \mathbb{R}^p$, $x_3 \in \mathbb{R}^n$ and $x_4 \in \mathbb{R}^{(m-1)q}$ are Lagrange multipliers, and $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}}, x_3^{\mathrm{T}}, x_4^{\mathrm{T}})^{\mathrm{T}}$. The iterative scheme of the sGS-ADMM algorithm for solving (2.8) is similar to the scheme in the context of DFMR as described in Table 1. The objective functions of the D, w, t subproblems are the same as those for DFMR, so the D, w, t subproblems have the same solutions as for the DFMR. Now we look at the remaining subproblems. The objective functions are strongly convex and smooth for the u-subproblem and v-subproblem. Their closed-form solutions can be obtained directly from the derivative of the augmented Lagrangian function (see Appendix A.4).

Finally, after k iterations, the s-subproblem can be written as

$$s^{k+1} = \arg\min_{s}\{\sum_{i=1}^{n}(1-s_i)\log(1-s_i)+s_i\log s_i - \langle x_3^k, s\rangle + \frac{\sigma}{2}\|u^{k+\frac{1}{2}} + s - y\|^2\}.$$

Unfortunately, there is no closed-form solution to the s-subproblem. But its objective function is smooth. We take the derivative of the objective function with respect to each component of s. The derivative with respect to the i-th component $(i = 1, 2, \cdots, n)$ is $\log s_i - \log(1 - s_i) + \sigma s_i + \sigma u_i^{k+\frac{1}{2}} - \sigma y_i - (x_3^k)_i$, where $(x_3^k)_i$ denotes the i-th component of x_3^k. Then we use the Bisection method to find the solution. More details are included in Appendix A.4.

2.2.3. Convergence analysis

In order to explore the convergence result, we denote $h(\alpha) = \sum_{i=1}^n (1-s_i)\log(1-s_i)+s_i \log s_i+P^*(w)+\delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)})+\delta(D; B_{\|\cdot\|_2(0;\lambda_1)})$, then there exists a self-adjoint and positive semi-definite linear operator Σ_α such that for all α, α', $\zeta \in \partial h(\alpha)$ and $\zeta' \in \partial h(\alpha')$, $\langle \zeta' - \zeta, \alpha' - \alpha\rangle \geq \|\alpha' - \alpha\|_{\Sigma_\alpha}^2$. We define $\Phi^*(u, v, \alpha, x) = (\mathbb{X}, \mathbb{Z}, I, 0)^{\mathrm{T}} u + (C, 0, 0, I)^{\mathrm{T}} v - Diag(I_{mq}, I_p, -I_n, -I_{(m-1)q})\alpha - (0, 0, y, 0)^{\mathrm{T}}$ and $\mathcal{M} := Diag(I, \sigma(I + CC^{\mathrm{T}}), \sigma\Sigma_I + \Sigma_\alpha, (\tau\sigma)^{-1} I_x) + s_\tau \sigma \Phi\Phi^*$, where the matrix $\Sigma_I = I_x = Diag(I_{mq}, I_p, I_n, I_{(m-1)q})$. Note that the operator $\mathcal{M}$ is different from that in DFMR. Let $\bar{\Omega}$ denote the optimal solution set satisfying the KKT condition.

Theorem 2.3. *The sequence $\{u^k, v^k, \alpha^k, x^k\}$ converges to the optimal solution $(\bar{u}, \bar{v}, \bar{\alpha}, \bar{x})$, where $(\bar{u}, \bar{v}, \bar{\alpha})$ is an optimal solution of (2.8), and $\bar{x}$ is an optimal solution of (2.6). Moreover, there exists $0 < \mu < 1$ such that for all $k \geq 1$, $dist_{\mathcal{M}}^2(\theta^{k+1}, \bar{\Omega}) \leq \mu dist_{\mathcal{M}}^2(\theta^k, \bar{\Omega})$.*

3. Consistency

3.1. DFMR

In this section, we investigate the consistency of $\hat{B}$ and $\hat{\gamma}$. Let B^* and γ^* be the true value of B and γ. We state the following conditions.
Condition 1. There exist two positive constants $\underline{C}_X$ and $\overline{C}_X$ such that

$$\underline{C}_X ||B||_F^2 \leq \frac{1}{n}\sum_{i=1}^n \langle X_i, B\rangle^2 \leq \overline{C}_X ||B||_F^2.$$

Condition 2. There exist two positive constants $\underline{C}_Z$ and $\overline{C}_Z$ that bound all eigenvalues of $n^{-1}\mathbb{Z}^T\mathbb{Z}$ from below and above, respectively.
Condition 3. The error ε_i satisfies $E\varepsilon_i = 0$ and follows the sub-Gaussian distribution, i.e., there exist two constants k and σ_0 such that $k^2[E(e^{|\varepsilon_i|^2/k^2}) - 1] \leq \sigma_0^2$.
Condition 4. Denote $r^* = \mathrm{rank}(B^*), s_1^* = \|C\mathrm{vec}(B^*)\|_0$, $s_2^* = \|\gamma^*\|_0$ and $s_3^* = \|A_p\gamma^*\|_0$. Take the tuning parameters as

$$\lambda_1 \leq \sqrt{n}/(r^*\|B^*\|_2),\ \lambda_2 \leq \sqrt{n}/(s_1^*\|C\mathrm{vec}(B^*)\|_\infty),$$
$$\lambda_3 \leq \sqrt{n}/(s_2^*\|\gamma^*\|_\infty),\ \lambda_4 \leq \sqrt{n}/(s_3^*\|A_p\gamma^*\|_\infty).$$

Condition 1 is the matrix version of Restricted Isometry condition which was suggested by [39]. It is widely used in the analysis for high-dimensional low-rank matrices, for example [24, 25, 42]. It indicates that the smallest eigenvalue of $n^{-1}\mathbb{X}^T\mathbb{X}$ is lower bounded by $\underline{C}_X$, and its largest eigenvalue is upper bounded by $\overline{C}_X$. Conditions 2-3 are commonly used in variable selection in high-dimensional linear regression, for example [2, 9, 54]. Theorem 3.1 establishes the consistency of the estimator matrix $\hat{B}$ and vector $\hat{\gamma}$.

Theorem 3.1. *Assume that Conditions 1-4 hold. Suppose that*

$$||\frac{1}{n}\sum_{i=1}^{n} z_i \otimes X_i||_F \leq \frac{1}{2}\min\{\underline{C}_X, \underline{C}_Z\}, \tag{3.1}$$

where the symbol $\otimes$ indicates Kronecker product. Then,

$$\begin{aligned} ||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2 \leq 16(k^2 + \sigma_0^2)\frac{\overline{C}_X mq + \overline{C}_Z p}{n\min^2\{\underline{C}_X, \underline{C}_Z\}} \\ + \frac{8(\lambda_1 r^* \|B^*\|_2 + \lambda_2 s_1^* \|C\text{vec}(B^*)\|_\infty + \lambda_3 s_2^* ||\gamma^*||_\infty + \lambda_4 s_3^* ||A_p\gamma^*||_\infty)}{n\min\{\underline{C}_X, \underline{C}_Z\}} \end{aligned}$$

with probability at least $1 - e^{-n\sigma_0^2/k^2} - c_1 e^{-c_2 mq} - c_3 e^{-c_4 p}$, where c_i, $i = 1, 2, 3, 4$, are positive constants. Further, if

$$\frac{\max\{mq, p\}}{n} \to 0 \tag{3.2}$$

as $n \to \infty$, it follows that $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2$ converges in probability to zero.

The assumption (3.1) assumes that X and z are only weakly correlated. Similar condition is used to study the consistency of bridge estimator for sparse regression with quadratic measurements in [11]. From Theorem 3.1, we learn that $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2$ converges at the rate of $O(\max\{mq, p\}/n)$.

3.2. DFMLR

Now we discuss the consistency of $\hat{B}$ and $\hat{\gamma}$ for DFMLR. Let B^* and γ^* be the true value of B and γ. Let $K_i = e^{\langle X_i, B^*\rangle + \langle z_i, \gamma^*\rangle}/(1 + e^{\langle X_i, B^*\rangle + \langle z_i, \gamma^*\rangle})^2$, $i = 1, 2, \ldots, n$.

Condition 5. There exist two positive constants $\underline{C}_X$ and $\overline{C}_X$ such that

$$\underline{C}_X ||B||_F^2 \leq \frac{1}{n}\sum_{i=1}^{n} K_i \langle X_i, B\rangle^2 \leq \overline{C}_X ||B||_F^2.$$

Condition 6. There exist two positive constants $\underline{C}_Z$ and $\overline{C}_Z$ that bound all eigenvalues of $n^{-1}\sum_{i=1}^{n} K_i z_i z_i^T$ from below and above, respectively.

Condition 7. Suppose that $\{\text{vec}(X_i)\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random vectors with mean 0, $k^2[E(e^{|\text{vec}(X_i)|^2/k^2}) - 1] \leq \sigma_0^2$ and $k^2[E(e^{|z_i|^2/k^2}) -$

$1] \leq \sigma_0^2$, where $k = \max\{||\text{vec}(X_i)||_{\psi_2}^2, ||z_i||_{\psi_2}^2\} < \infty$.

Condition 8. Denote $r^* = \text{rank}(B^*), s_1^* = \|C\text{vec}(B^*)\|_0,\ s_2^* = \|\gamma^*\|_0$ and $s_3^* = \|A_p\gamma^*\|_0$. Take the tuning parameters

$$\lambda_1 \leq \sqrt{n}/(r^*\|B^*\|_2),\ \lambda_2 \leq \sqrt{n}/(s_1^*\|C\text{vec}(B^*)\|_\infty),$$
$$\lambda_3 \leq \sqrt{n}/(s_2^*\|\gamma^*||_\infty),\ \lambda_4 \leq \sqrt{n}/(s_3^*||A_p\gamma^*||_\infty).$$

Condition 7 assumes that $\text{vec}(X)$ and z both follow sub-Gaussian distributions, which is also assumed by [10] in high-dimensional trace regression with a nuclear norm penalty. Theorem 3.2 provides the convergence rate and consistency results for $\hat{B}$ and $\hat{\gamma}$.

Theorem 3.2. *Assume that Conditions 5-8 hold. Suppose that*

$$||\frac{1}{n}\sum_{i=1}^{n} K_i z_i \otimes X_i||_F \leq \frac{1}{2}\min\{\underline{C}_X, \underline{C}_Z\}, \tag{3.3}$$

then

$$||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2 \leq \frac{16k^2(mq+p)}{n\min^2\{\underline{C}_X, \underline{C}_Z\}}$$
$$+\frac{4(\lambda_1 r^*\|B^*\|_2 + \lambda_2 s_1^*||C\text{vec}(B^*)||_\infty + \lambda_3 s_2^*\|\gamma^*||_\infty + \lambda_4 s_3^*\|A_p\gamma^*||_\infty)}{n\min\{\underline{C}_X, \underline{C}_Z\}}$$

with probability at least $1 - c_1 e^{-c_2 mq} - c_3 e^{-c_4 p}$, *where* c_i, $i = 1, 2, 3, 4$, *are positive constants. Further, if*

$$\frac{\max\{mq, p\}}{n} \to 0$$

as $n \to \infty$, *it follows that* $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2$ *converges in probability to zero.*

Theorem 3.2 indicates that $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2$ also converges at the rate of $O(\max\{mq, p\}/n)$ under the context of DFMLR.

4. Simulation

In this section, we demonstrate the performance of DFMR and DFMLR with numerical experiments. To evaluate estimation performance, we computed the average root mean squared errors (RMSEs) for each estimator of B and γ, denoted by RMSE(B) and RMSE(γ) based on 100 repetitions. To evaluate the prediction performance, we use a testing dataset with the same sample size as the training dataset. For the DFMR estimator, the prediction error is the root mean squared error over the testing dataset, denoted by RMSE(y). Specifically, $\text{RMSE}(y) = [n^{-1}\sum_{i=1}^n (y_i - \hat{y}_i)^2]^{1/2}$,where $\hat{y}_i$ is the fitted value for the ith observation. For the DFMLR estimator, the prediction error is the misclassification rate, denoted by MCR, since we consider the prediction as classification [23] on the testing dataset. The experiments were implemented in MATLAB 2017b on a desktop computer with i5-8250U, 1.80GHZ CPU and 8 GB of RAM.

4.1. DFMR

We compare the following three estimators: the DFMR estimator obtained by solving (2.5) and two estimators obtained by two regularization methods from [50], i.e., matrix Lasso regression estimator (MLR) obtained from

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

and Lasso regression estimator (LR) from solving

$$\min_{B,\gamma} \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle X_i, B\rangle - \langle z_i, \gamma\rangle)^2 + \lambda\|\mathrm{vec}(B)\|_1.$$

The MLR and LR estimators are computed by the MATLAB toolbox TensorReg from [50]. For the DFMR estimator, we adopt the common choice for tuning parameters as

$$\lambda_1 = \alpha_1\|\mathbb{X}^T y\|_\infty,\ \lambda_2 = \alpha_2\lambda_1,\ \lambda_3 = \alpha_3\|\mathbb{Z}^T y\|_\infty,\ \lambda_4 = \alpha_4\lambda_3,$$

where $0 < \alpha_1, \alpha_3 < 1, \alpha_2, \alpha_4 > 0$.

We fixed $m = 100, q = 50$, $p = 250$ and varied n from 200 to 1600. Each element in the matrix-valued predictor X and vector-valued predictor z was drawn independently from the standard normal distribution. We generated $B_0 = b_1 b_2^{\mathrm{T}}$, where $b_1 \in \mathbb{R}^{m\times 1}$ and $b_2 \in \mathbb{R}^{q\times 1}$ are two vectors. The first $m/4$ elements and last $m/4$ elements in b_1 were 0 and the middle $m/2$ elements were 1, and b_2 had similar structure. Then the elements in B_0 were 0 except that the elements in the center square were 1. Then rank of B_0 is 1. Let $R < \min(m, q)$ denote the rank of the matrix coefficient B. Then B was constructed by adding a $(R-1)\times(R-1)$ identity matrix above the left corner of the center square in B_0. Let s denote the sparsity level of γ, $0 \le s \le 1$, which means that the proportion of the nonzero elements in γ is s. The nonzero elements in γ all took value 1. We fixed $R = 5$ and $s = 0.01$. The error ε was drawn from a standard normal distribution. The estimation and prediction performance of each estimator is summarized in Table 2. The numbers in the parentheses are the standard deviations computed from the 100 repetitions.

The DFMR estimator has the best estimation and prediction performance for all sample sizes. For example, at sample size 400, the DFMR estimator reduces RMSE(B) and RMSE(γ) by 92% and 96% compared to the MLR estimator; and reduced RMSE(B) and RMSE(γ) by 93% and 97% compares to the LR estimator. For prediction performance, the DFMR estimator reduces RMSE(y) by 92% compared to the MLR estimator and 93% compared to the LR estimator. The MLR estimator has very large RMSEs when the sample size is small, but it performs much better when the sample size increases. The LR estimator also has a large RMSE in estimation and prediction, but in contrast to the MLR estimator, its performance does not improve much when the sample size increases.

Table 2

Comparison of the DFMR, MLR and LR estimators with different sample sizes

	RMSE(B)			RMSE(γ)			RMSE(y)		
n	DFMR	MLR	LR	DFMR	MLR	LR	DFMR	MLR	LR
200	0.06 (0.0015)	0.37 (0.0003)	0.38 (0.0005)	0.11 (0.0002)	9.04 (0.1210)	3.43 (0.0176)	0.06 (0.0019)	0.70 (0.0078)	0.46 (0.0010)
400	0.03 (0.0007)	0.36 (0.0004)	0.42 (0.0007)	0.07 (0.0042)	1.96 (0.0115)	2.41 (0.0129)	0.03 (0.0009)	0.38 (0.0007)	0.44 (0.0015)
800	0.03 (0.0006)	0.06 (0.0013)	0.41 (0.0006)	0.05 (0.0038)	0.18 (0.0044)	1.35 (0.0060)	0.03 (0.0008)	0.06 (0.0011)	0.44 (0.0013)
1200	0.02 (0.0005)	0.04 (0.0007)	0.39 (0.0006)	0.03 (0.0023)	0.09 (0.0035)	0.89 (0.0050)	0.02 (0.0006)	0.04 (0.0007)	0.39 (0.0011)
1600	0.01 (0.0004)	0.03 (0.0004)	0.35 (0.0007)	0.02 (0.0011)	0.06 (0.0014)	0.66 (0.0045)	0.01 (0.0004)	0.03 (0.0004)	0.35 (0.0010)

Table 3

Comparison of the DFMR, MLR and LR estimators with different dimension of γ

	RMSE(B)			RMSE(γ)			RMSE(y)		
p	DFMR	MLR	LR	DFMR	MLR	LR	DFMR	MLR	LR
100	0.03 (0.0007)	0.33 (0.0007)	0.37 (0.0006)	0.07 (0.0072)	1.46 (0.0120)	2.04 (0.0161)	0.03 (0.0008)	0.34 (0.0010)	0.42 (0.0018)
150	0.04 (0.0008)	0.37 (0.0005)	0.39 (0.0008)	0.09 (0.0070)	2.04 (0.0103)	2.53 (0.0178)	0.04 (0.0010)	0.36 (0.0016)	0.44 (0.0021)
200	0.04 (0.0011)	0.38 (0.0003)	0.40 (0.0006)	0.10 (0.0002)	2.57 (0.0128)	2.82 (0.0184)	0.05 (0.0012)	0.40 (0.0016)	0.45 (0.0017)
250	0.05 (0.0011)	0.38 (0.0002)	0.41 (0.0009)	0.11 (0.0002)	3.53 (0.0182)	4.27 (0.0280)	0.05 (0.0012)	0.41 (0.0008)	0.46 (0.0018)

Table 4

Comparison of the DFMR, MLR and LR estimators with different dimensions of B

		RMSE(B)			RMSE(γ)			RMSE(y)		
m	q	DFMR	MLR	LR	DFMR	MLR	LR	DFMR	MLR	LR
50	50	0.05 (0.0012)	0.36 (0.0006)	0.40 (0.0010)	0.07 (0.0041)	2.82 (0.0239)	2.97 (0.0236)	0.05 (0.0014)	0.45 (0.0012)	0.46 (0.0018)
100	50	0.05 (0.0011)	0.38 (0.0002)	0.41 (0.0009)	0.11 (0.0002)	3.53 (0.0182)	4.27 (0.0280)	0.05 (0.0012)	0.41 (0.0008)	0.46 (0.0018)
100	100	0.08 (0.0012)	0.39 (0.0001)	0.43 (0.0005)	0.13 (0.0035)	6.08 (0.0259)	6.16 (0.0283)	0.07 (0.0013)	0.50 (0.0012)	0.49 (0.0018)
200	100	0.10 (0.0008)	0.42 (0.0006)	0.45 (0.0003)	0.15 (0.0015)	6.23 (0.0146)	7.29 (0.0258)	0.09 (0.0006)	0.52 (0.0007)	0.57 (0.0015)

Then we fixed $n = 300$, $m = 100$, $q = 50$ and varied the dimension of γ under the setting of Table 2. The results are summarized in Table 3. The DFMR estimator still has the best estimation and prediction performance. Both MLR and LR have very large RMSE in the estimation of γ, because their objective functions do not consider the sparsity structure on γ.

We also fixed $n = 300$, $p = 250$ and varied the dimension of B under the setting of Table 2. The results are presented in Table 4. DFMR again shows the best estimation and prediction performance for different m and q.

4.2. DFMLR

In this section, we compared the numerical performance of three estimators: the DFMLR estimator obtained by solving (2.8), and two estimators from two regularization methods in [50], i.e., matrix Lasso penalized logistic estimator (denoted by MLLR) from solving

$$\min_{B,\gamma} \sum_{i=1}^{n} \log(1 + e^{\langle X_i, B\rangle + \langle z_i, \gamma\rangle}) - y_i(\langle X_i, B\rangle + \langle z_i, \gamma\rangle) + \lambda\|B\|_*,$$

Table 5
Comparison of the DFMLR, MLLR and LLR estimators with different sample sizes

	RMSE(B)			RMSE(γ)			MCR		
n	DFMLR	MLLR	LLR	DFMLR	MLLR	LLR	DFMLR	MLLR	LLR
200	0.21 (0.0058)	0.39 (0.0005)	0.39 (0.0003)	0.18 (0.0303)	2.52 (0.1814)	5.99 (0.4406)	0.18 (0.0191)	0.53 (0.0151)	0.53 (0.0158)
400	0.18 (0.0038)	0.39 (0.0002)	0.39 (0.0001)	0.16 (0.0239)	2.15 (0.1132)	2.06 (0.0161)	0.15 (0.0105)	0.44 (0.0169)	0.51 (0.0147)
800	0.17 (0.0026)	0.37 (0.0005)	0.38 (0.0004)	0.14 (0.0167)	0.30 (0.0122)	0.39 (0.0118)	0.15 (0.0084)	0.43 (0.0075)	0.48 (0.0107)
1200	0.12 (0.0049)	0.35 (0.0004)	0.38 (0.0004)	0.12 (0.0064)	0.20 (0.0057)	0.26 (0.0087)	0.08 (0.0038)	0.29 (0.0094)	0.45 (0.0105)
1600	0.10 (0.0024)	0.34 (0.0004)	– –	0.01 (0.0112)	0.14 (0.0052)	– –	0.07 (0.0048)	0.23 (0.0099)	– –

and Lasso penalized logistic estimator (denoted by LLR) from solving

$$\min_{B,\gamma} \sum_{i=1}^{n} \log(1 + e^{\langle X_i, B\rangle + \langle z_i, \gamma\rangle}) - y_i(\langle X_i, B\rangle + \langle z_i, \gamma\rangle) + \lambda\|\mathrm{vec}(B)\|_1.$$

The MLLR and LLR estimators are computed by the MATLAB toolbox TensorReg from [50]. For the DFMLR estimator, we choose λ_1, λ_2, λ_3 and λ_4 using [18]:

$$\lambda_1 = \alpha_1\|\mathbb{X}^T y_0\|_\infty,\ \lambda_2 = \alpha_2\lambda_1,\ \lambda_3 = \alpha_3\|\mathbb{Z}^T y_0\|_\infty,\ \lambda_4 = \alpha_4\lambda_3,$$

where $0 < \alpha_1, \alpha_3 < 1, \alpha_2, \alpha_4 > 0$, for $i = 1, 2, \cdots, n$, $y_0(i) = -\frac{n_-}{n}$ if $y_i = 1$, $y_0(i) = -\frac{n_+}{n}$ if $y_i = 0$, n_+ is the number of y_i's that takes value 1, and $n_- = n - n_+$ is the number of y_i's that takes value 0.

We fixed $m = 100$, $q = 50$, $p = 250$, $R = 5$ and $s = 0.01$ and varied n from 200 to 1600. The coefficients B and γ, the matrix-valued predictor X and the vector-valued predictor z were generated in the same way as in Section 4.1. The measures of estimation performance RMSE(B) and RMSE(γ) as well as the misclassification rate MCR of the DFMLR, MLLR and LLR estimators are summarized in Table 5. The numbers in the parentheses are the standard deviations computed from the 100 repetitions.

Based on the results in Table 5, the DFMLR estimator has the best estimation and classification performance for all sample sizes. Take the sample size 400 as an example, the DFMLR estimator reduces RMSE(B) and RMSE(γ) by 54% and 93% compared to the MLLR estimator. It reduces RMSE(B) by 54% and reduces RMSE(γ) by 92% compared to the LLR estimator. For the prediction performance, the DFMLR reduces the misclassification rate by 66% compared to the MLLR estimator and 71% compared to the LLR estimator. When the sample size is 1600, the MATLAB program for computing the LLR estimator fails, and thus no results are recorded.

We then fixed the sample size n at 300, and varied the dimension of γ under the setting that produced Table 5. Finally we fixed $n = 300$ and $p = 250$, but varied the dimension of B. The results are shown in Tables 6 and 7. In both settings, the DFMLR estimator has the smallest RMSEs for estimation of B and γ and smallest misclassification rate compared to other estimators.

TABLE 6

Comparison of the DFMLR, MLLR and LLR estimators with different dimension of γ

	RMSE(B)			RMSE(γ)			MCR		
p	DFMLR	MLLR	LLR	DFMLR	MLLR	LLR	DFMLR	MLLR	LLR
100	0.12 (0.0074)	0.36 (0.0003)	0.37 (0.0005)	0.10 (0.0072)	0.87 (0.0433)	0.69 (0.0343)	0.09 (0.0113)	0.43 (0.0179)	0.45 (0.0188)
150	0.14 (0.0103)	0.37 (0.0008)	0.38 (0.0013)	0.12 (0.0056)	1.40 (1.0570)	2.73 (5.7965)	0.10 (0.0140)	0.43 (0.0156)	0.48 (0.0174)
200	0.17 (0.0055)	0.38 (0.0001)	0.38 (0.0002)	0.13 (0.0277)	1.45 (0.0990)	2.80 (0.1723)	0.14 (0.0135)	0.45 (0.0147)	0.50 (0.0164)
250	0.21 (0.0036)	0.39 (0.0002)	0.39 (0.0004)	0.17 (0.0261)	1.52 (0.0507)	2.84 (0.0560)	0.18 (0.0118)	0.47 (0.0150)	0.53 (0.0144)

TABLE 7

Comparison of the DFMLR, MLLR and LLR estimators with different dimensions of B

		RMSE(B)			RMSE(γ)			MCR		
m	q	DFMLR	MLLR	LLR	DFMLR	MLLR	LLR	DFMLR	MLLR	LLR
50	50	0.17 (0.0093)	0.38 (0.0006)	0.38 (0.0003)	0.12 (0.0104)	1.36 (0.1335)	2.65 (0.1172)	0.13 (0.0154)	0.45 (0.0160)	0.52 (0.0167)
100	50	0.21 (0.0036)	0.39 (0.0002)	0.39 (0.0004)	0.17 (0.0261)	1.52 (0.0507)	2.84 (0.0560)	0.18 (0.0118)	0.47 (0.0150)	0.53 (0.0144)
100	100	0.25 (0.0025)	0.41 (0.0001)	0.41 (0.0005)	0.17 (0.0312)	1.60 (0.0624)	2.90 (0.0802)	0.21 (0.0123)	0.51 (0.0126)	0.55 (0.0154)
200	100	0.29 (0.0023)	0.42 (0.0003)	0.42 (0.0001)	0.18 (0.0175)	1.84 (0.0542)	3.11 (0.0492)	0.23 (0.0130)	0.54 (0.0147)	0.56 (0.0129)

4.3. Signal shapes

In this section, we illustrate the effect of the nuclear norm penalty in estimation. For this purpose, we introduce a new estimator, called the fused matrix Lasso regression estimator, denoted by FMR. The FMR estimator is obtained from the following optimization problem, whose objective function is the same as that in (1.3), but without the nuclear norm penalty:

$$\min_{B,\gamma} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \langle X_i, B\rangle - \langle z_i, \gamma\rangle\right)^2 + \lambda_2 \sum_{j=2}^{m} \|B_{j\cdot} - B_{(j-1)\cdot}\|_1 + \lambda_3\|\gamma\|_1 + \lambda_4 \sum_{k=2}^{p} |\gamma_k - \gamma_{k-1}|.$$

Thus the effect of the nuclear norm penalty can be illustrated by the comparison between the DFMR estimator and the FMR estimator. We generated the data from (1.1) without the vector-valued predictor z, i.e., $y = \langle X, B\rangle + \varepsilon$, where ε followed a standard normal distribution. The matrix predictor X was a 64×64 matrix with entries independently drawn from the standard normal distribution. The elements in the coefficient matrix B were binary, which took value 1 according to a variety of signal shapes displayed in the first column of Figure 1. The sample size n was fixed at 500. The DFMR, FMR, MLR, and LR estimators were plotted in the second, third, fourth and fifth columns of Figure 1, and their RMSE(B) are arranged in Table 8.

The rank of the true signals varies from 9 to 20. Regardless of the rank, the DFMR estimator has the best visual recovery of the true signal, followed by the FMR estimator (with only the fused Lasso penalty) and the MLR estimator

TABLE 8
Comparison of DFMR, FMR, MLR and LR on RMSE(B)

Estimator	Shape				
	Cross	Star	Hook	Windwill	Mickey
DFMR	0.07 (0.0010)	0.09 (0.0011)	0.09 (0.0016)	0.06 (0.0014)	0.07 (0.0009)
FMR	0.14 (0.0026)	0.17 (0.0024)	0.15 (0.0034)	0.12 (0.0030)	0.11 (0.0025)
MLR	0.13 (0.0007)	0.15 (0.0005)	0.18 (0.0005)	0.12 (0.0004)	0.11 (0.0006)
LR	0.21 (0.0013)	0.24 (0.0010)	0.24 (0.0010)	0.10 (0.0020)	0.14 (0.0015)

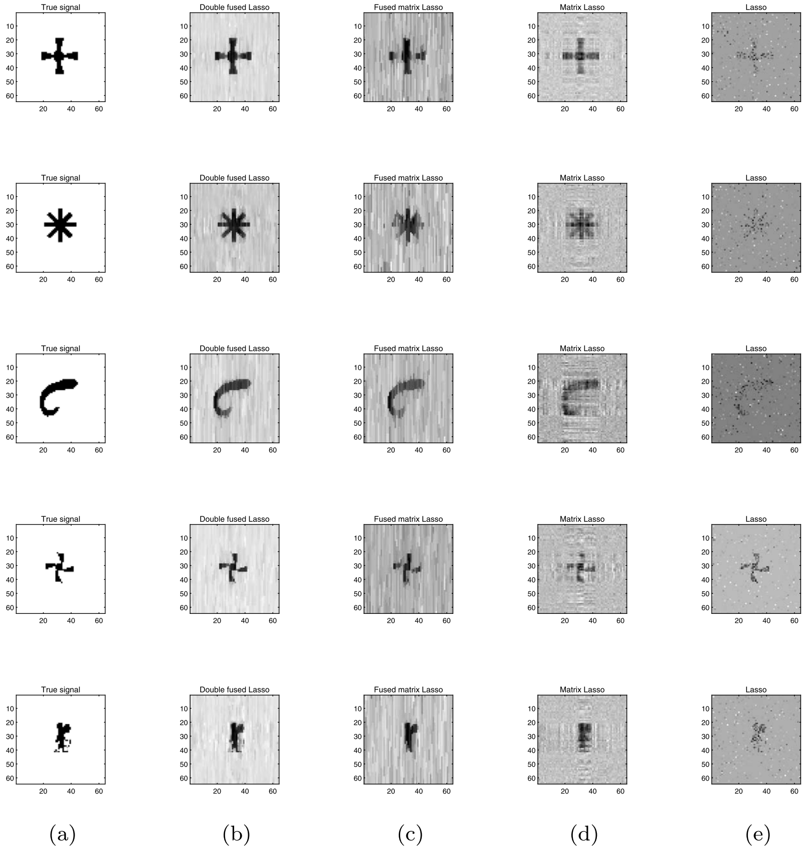

FIG 1. *(a) True signal, (b) DFMR estimator, (c) FMR estimator, (d) MLR estimator, (e) LR estimator.*

(with only the nuclear norm penalty). The LR estimator (with the L_1 penalty, but no nuclear norm or fused Lasso penalties) has the worst visual recovery. This trend is confirmed by the RMSEs in Table 8. The DFMR estimator has the smallest RMSE(B), followed by the FMR and MLR estimators, and with the LR estimator trailing behind. Take the shape of Cross as an example, the DFMR estimator reduces RMSE(B) by 50% compared to the FMR estimator, 46% compared to MLR estimator and 67% compared to the LR estimator. Clearly, the nuclear norm penalty is helpful to explore the structure of the coefficient matrix B.

5. Examples

5.1. Bike sharing dataset

The bike sharing system plays a more and more important role in urban traffic, due to its positive effects in energy consumption, environmental protection and public health. Currently, there are nearly 900 bike-sharing programs worldwide operating about 1,000,000 bicycles. The rental and return of bicycles has become quite convenient for users. Through automated stations, users can rent a bike from one position and return at a different position. Bike-sharing demand is highly dependent on weather conditions and social factors such as temperature, precipitation, and holidays. The bike sharing dataset [13] consists of bike rental records from Capital Bikeshare, the metro Washington D.C.'s bike share system, for a two-year period from January 1st, 2011 to December 31th, 2012. The weather conditions, including weather type (sunny, mist or others), temperature, apparent temperature, humidity and wind speed are measured every hour. We took the measurements as a 24×6 matrix-valued predictor X. The vector-valued predictor z contains indicators of months (January to December), days in a week (Sunday to Saturday), year (2011 or 2012) and holiday (work day or not). The response y is the daily aggregated count of rented bikes. The DFMR, MLR and LR estimators were computed. Due to the Lasso and fused Lasso penalties on γ, the DFMR estimator shows that Mondays have the least demand for bike rental; then the demand increases on Tuesdays and Wednesdays, where Wednesdays reach the same level of Sundays, the demand further increases on Thursdays and reaches its peak on Fridays and Saturdays, then it falls on Sundays. The MLR estimator and the LR estimator show the same weekly pattern but without a sparse pattern. For example, the coefficients for Fridays and Saturdays are very similar under the MLR and LR, while they are the same under the DFMR. Due to the fused Lasso penalty on B, DFMR estimator reveals that the coefficients for temperatures are very similar before 9am or after 7pm, indicating that the time of the day has an effect on the rental demand. Without the fused Lasso penalty, the MLR estimator of B is variant and does not have an obvious pattern. Without both the fused Lasso penalty and the nuclear norm penalty, the LR estimator of B is zero.

The prediction performance measured by the average RMSE(y) was computed by 5-fold or 10-fold cross-validation (CV) with 100 random splits. The results are in Table 9. The numbers in the parentheses are standard deviations of the RMSE(y). The DFMR estimator again has smallest prediction error. Take 5-fold CV as an example, it reduces the RMSE(y) by 36.5% and 28.1% compared to the MLR estimator and LR estimator.

Table 9
Comparison of the DFMR, MLR and LLR estimators

	DFMR	MLR	LR
5-fold CV	7.62e+02 (7.50e+00)	1.20e+03 (2.65e+01)	1.06e+03 (6.26e+00)
10-fold CV	7.57e+02 (4.73e+00)	1.18e+03 (2.44e+01)	1.05e+03 (4.86e+00)

Table 10
Comparison of the DFMLR, MLLR and LLR estimators on misclassification rate

	DFMLR	MLLR	LLR
5-fold CV	0.0812 (0.0043)	0.1657 (0.0064)	0.1648 (0.0059)
10-fold CV	0.0809 (0.0042)	0.1678 (0.0055)	0.1663 (0.0052)

5.2. Diabetes dataset

The physical examination information for 2476 staffs at the Beijing Jiaotong University is collected by the university clinic from 2016 to 2018. Out of the 2476 staffs, 237 staffs were diagnosed to have diabetes in 2018. During the physical exam each year, a total of 62 measurements are recorded for each patient including blood sugar concentration, dietary preferences, concentration and volume of erythrocyte, leukocyte and platelets, and facial features, which yields a 62×3 matrix. In addition, information on seven characteristics that are relatively stable over the three-year period including gender, education, occupation, or disability status is available for each staff. We took the characteristics as the vector-valued predictor and physical exam results as the matrix-valued predictor. The response y_i is a binary variable which takes value 1 if the patient is diabetic.

The misclassification rates of the DFMLR, MLLR and LLR estimators were computed by 5-fold or 10-fold CV with 100 random splits. The results are included in Table 10. The numbers in the parentheses are standard deviations. For both 5-fold and 10-fold CV, the DFMLR estimator is able to reduce the misclassification rate by more than 50% compared to the LLR estimator or the MLLR estimator.

5.3. COVID-19 dataset

The COVID-19 dataset [47] consists of daily measurements related to COVID-19 for 138 countries around the world. We focus on the period June 13, 2020 to July 12, 2020. The response y is the total count of newly confirmed case in this 30-day period for each country. The matrix predictor X is the COVID-19 related government policy every day. The policies include school-closing, restrictions on gathering, stay-at-home requirement, income support and so on. Each of these policies may have several levels, for example, school closing includes no closing, recommend closing, require some closing (e.g. just high school) or

TABLE 11
Comparison of the DFMR, MLR and LLR estimators

	DFMR	MLR	LR
5-fold CV	0.7445 (0.0741)	1.5592 (0.4240)	1.0713 (0.1760)
10-fold CV	0.6360 (0.0479)	1.1939 (0.2384)	0.8849 (0.0895)

require all closing, which varied during the 30-day period. There are a total of 38 measurements from government policies, thus the dimension of X is 30×38. The vector-valued predictor z contains 23 characteristics of each country that remain constant or relatively steady during the 30-day period, for example, male and female population, gross domestic product (GDP), diabetes prevalence (percentage of persons with diabetes in population), number of nurses and smoking prevalence (percentage of smokers). The sample size is $n = 138$. The average RMSE(y) for the DFMR, LMR and LR estimators computed from CV are summarized in Table 11. The DFMR estimator has the smallest predictor error. Take 5-fold CV as an example, it reduces RMSE(y) by 52.3% compared to the MLR estimator and 30.5% compared to the LR estimator.

6. Concluding remarks

In this paper, we propose a regularized method in linear regression and logistic regression which can incorporate high-dimensional matrix-valued predictor and vector-valued predictor. The proposed method can be extended to models with tensor-valued predictors, which has many applications in neuroimaging and signal processing fields. In addition, the proposed method can be adapted to other generalized linear regression model such as Poisson regression, which is widely used in medical insurance, business statistics and geography [6].

Acknowledgments

The authors would like to thank the editor and anonymous referees for their invaluable comments and suggestions which are very helpful for improving the paper. The authors thank Professor Defeng Sun from Hong Kong Polytechnic University for his constructive comments and encouragements.

Appendix A: Appendix

A.1. Moreau-envelope function and proximal mapping

We present Moreau envelope function and proximal mapping. In particular, we list the explicit form of proximal mapping for specific functions, such as the indictor function, L_1-norm regularization function, fused Lasso regularization function, nuclear norm regularization function and matrix indictor function.

Let $p : \mathbb{R}^n \to (-\infty, +\infty]$ be a closed proper convex function such that for a given $\nu > 0$, the Moreau envelope function $\psi_p(\cdot)$ of p [32] is defined by

$$\psi_{p/\nu}(x) = \min_{z\in\mathbb{R}^n} \left\{ p(z) + \frac{\nu}{2}\|z - x\|^2 \right\}, \quad \forall\, x \in \mathbb{R}^n, \tag{A.1}$$

and the corresponding solution is called as the proximal mapping:

$$Prox_{p/\nu}(x) = \arg\min_{z\in\mathbb{R}^n} \left\{ p(z) + \frac{\nu}{2}\|z - x\|^2 \right\}, \quad \forall\, x \in \mathbb{R}^n.$$

Let $p : \mathbb{R}^n \to (-\infty, +\infty]$ be a closed proper convex function, then the Fenchel conjugate function of p is defined as $p^*(x) := \sup\limits_{x'\in\mathbb{R}^n} \left\{ \langle x, x' \rangle - p(x') \right\}, \forall x \in \mathbb{R}^n$.

Proposition A.1. *[33] Let $p : \mathbb{R}^n \to (-\infty, +\infty]$ be a closed proper convex function, and $p^*(x)$ be its Fenchel conjugate function. Then for any $t > 0$,*

$$Prox_{tp}(x) + tProx_{p^*/t}(x/t) = x, \quad \forall x \in \mathbb{R}^n. \tag{A.2}$$

The equality (A.2) is often referred to as the Moreau identity.

Now we discuss the proximal mapping of problem (A.1) with $\nu = 1$. We can obtain the explicit form of proximal mapping for some special functions. For example, if $p(z) = \delta(z; \Omega)$, where Ω is a nonempty closed convex set. The proximal mapping of the indicator function δ_Ω is the projection operator on the set Ω:

$$\begin{aligned} Prox_{\delta_\Omega}(x) &= \arg\min_{z\in\mathbb{R}^n} \left\{ \delta(z; \Omega) + \frac{1}{2}\|z - x\|^2 \right\} \\ &= \arg\min_{z\in\Omega} \left\{ \|z - x\|^2 \right\} = \Pi(x; \Omega). \end{aligned}$$

If $\Omega = B_{\|\cdot\|_\infty(0;r)}$, the proximal mapping of $\delta(x; \Omega)$ is

$$Prox_{\delta_\Omega}(x) = \Pi(x; \Omega) = x - sign(x) \cdot \max\{|x| - r, 0\}. \tag{A.3}$$

If $p(z) = \lambda\|z\|_1$, the proximal mapping of p is $Prox_p(x) = shrink(x, \lambda) := sign(x) \cdot \max\left\{|x| - \lambda, 0\right\}$, which is called as the soft-thresholding operator in [16].

If $p(z) = \lambda_1\|z\|_1 + \lambda_2\|A_p z\|_1$, where $\lambda_1, \lambda_2 \geq 0$ are given parameters and $A_p \in \mathbb{R}^{(p-1)\times p}$,

$$A_p = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}, \tag{A.4}$$

then the proximal mapping of p is

$$Prox_p(x) = \arg\min_{z\in\mathbb{R}^n} \left\{ \lambda_1\|z\|_1 + \lambda_2\|A_p z\|_1 + \frac{1}{2}\|z - x\|^2 \right\}, \quad \forall\, x \in \mathbb{R}^n. \tag{A.5}$$

If $\lambda_1 = 0$ in (A.5), we denote the proximal mapping of $p(z) = \lambda_2\|A_p z\|_1$ by $z_{\lambda_2}(x)$, and $z_{\lambda_2}(x)$ is

$$z_{\lambda_2}(x) = \arg\min_{z\in\mathbb{R}^n} \left\{ \lambda_2\|A_p z\|_1 + \frac{1}{2}\|z-x\|^2 \right\}, \quad \forall x \in \mathbb{R}^n.$$

Proposition A.2. *[16] Let $p(z) = \lambda_1\|z\|_1 + \lambda_2\|A_p z\|_1$, $A_p \in \mathbb{R}^{(p-1)\times p}$ has the form as in (A.4), then we have*

$$Prox_p(x) = Prox_{\lambda_1\|\cdot\|_1}(z_{\lambda_2}(x)) = sign(z_{\lambda_2}(x)) \cdot max\{|z_{\lambda_2}(x)| - \lambda_1, 0\}. \tag{A.6}$$

Now we present the proximal mapping for the matrix-form function. Let $p(M) = \|M\|_*$, then

$$Prox_p(D) = \arg\min_{M\in\mathbb{R}^{m\times q}} \left\{ \|M\|_* + \frac{\nu}{2}\|M-D\|_F^2 \right\}, \quad \forall\, D \in \mathbb{R}^{m\times q},$$

and it has a closed-form solution, which is given by

$$Prox_p(D) = U_D Diag(\hat{\varsigma})V_D^{\mathrm{T}},\ \hat{\varsigma} = shrink(\varsigma, \frac{1}{\nu}) = sign(\varsigma)\cdot \max\{|\varsigma| - \frac{1}{\nu}, 0\},$$

where U_D, V_D, Σ_D are from the singular value decomposition of D, i.e., $D = U_D\Sigma_D V_D^{\mathrm{T}}$, and ς is a vector that contains the diagonal element of Σ_D. The proof can be found in [3].

Let $\nu > 0$, and $p(M) = \delta(M;\Omega^*)$, then the proximal mapping of p is

$$Prox_p(D) = \arg\min_{M\in\mathbb{R}^{m\times q}} \left\{ \delta(M;\Omega^*) + \frac{\nu}{2}\|M-D\|_F^2 \right\}, \quad \forall\, D \in \mathbb{R}^{m\times q}$$

with $\Omega^* = B_{\|\cdot\|_2(0;\lambda)}$ and it also has a closed-form solution, which is

$$Prox_p(D) = U_D Diag(\hat{\varsigma})V_D^{\mathrm{T}}, \quad \hat{\varsigma} = \Pi(\varsigma; B_{\|\cdot\|_\infty(0;\lambda)}). \tag{A.7}$$

In special cases, the proximal mapping is a projection and plays an important role in solving the problem. Based on it, we derive an efficient sGS-ADMM algorithm to solve DFMR and DFMLR.

A.2. An introduction to sGS-ADMM algorithm

Now we give a brief introduction on sGS technique and sGS-ADMM algorithm for a general convex composite programming model as discussed in [5].

The sGS means the symmetric Gauss Seidel method which is an extension of Gauss Seidel (GS) method. The GS method has been applied in linear or nonlinear systems, unconstrained or constrained optimization problems, see [19, 37, 46]. Chen et al. [5] designed the sGS method to solve convex composite conic programming. We now illustrate the GS and sGS methods with an example. Consider solving the linear system

$$Ax = b,$$

where $A \in R^{m \times n}$ is the coefficient matrix and $b \in R^m$ is the right-hand side.

GS: successively update the elements of x in a fixed order and turn to the first one once the last one is updated.

$$\boxed{x_1 \to x_2 \to \cdots \to x_n \to x_1} \to x_2 \to \cdots$$

sGS: successively update the elements of x in a fixed order and turn to the first one in reverse order once the last one is updated.

$$\boxed{x_1 \to x_2 \to \cdots \to x_n \to x_{n-1} \to x_{n-2} \to \cdots x_2 \to x_1} \to x_2 \to \cdots$$

Let m and n be two nonnegative integers, $\mathcal{X}, \mathcal{Y}_i, 1 \leq i \leq m$ and $\mathcal{Z}_j, 1 \leq j \leq n$, be finite dimensional Euclidean spaces. Define $\mathcal{Y} := \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m$ and $\mathcal{Z} := \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$. Consider the following general convex composite programming model:

$$\min_{y \in \mathcal{Y}, z \in \mathcal{Z}} \{p_1(y_1) + f(y_1, \cdots, y_m) + q_1(z_1) + g(z_1, \cdots, z_n) \mid \mathcal{A}^* y + \mathcal{B}^* z = c\} \text{(A.8)}$$

where $p_1 : \mathcal{Y}_1 \to (-\infty, +\infty]$ and $q_1 : \mathcal{Z}_1 \to (-\infty, +\infty]$ are two closed proper convex functions, $f : \mathcal{Y} \to (-\infty, +\infty)$ and $g : \mathcal{Z} \to (-\infty, +\infty)$ are continuously differentiable convex functions whose gradients are Lipschitz continuous. The linear mappings $\mathcal{A} : \mathcal{Y} \to \mathcal{X}$ and $\mathcal{B} : \mathcal{Z} \to \mathcal{X}$ are defined such that their adjoints are given by $\mathcal{A}^* y = \sum\limits_{i=1}^{m} \mathcal{A}_i^* y_i$ for $y = (y_1, \cdots, y_m) \in \mathcal{Y}$ and $\mathcal{B}^* z = \sum\limits_{j=1}^{n} \mathcal{B}_j^* z_j$ for $z = (z_1, \cdots, z_n) \in \mathcal{Z}$, where $\mathcal{A}_i^* : \mathcal{Y}_i \to \mathcal{X}$ and $\mathcal{B}_j^* : \mathcal{Z}_j \to \mathcal{X}$ are the adjoints of the linear mappings $\mathcal{A}_i : \mathcal{X} \to \mathcal{Y}_i$ and $\mathcal{B}_j : \mathcal{X} \to \mathcal{Z}_j$, respectively.

The augmented Lagrangian function of problem (A.8) is defined as follows.

$$\begin{aligned} \mathcal{L}_\sigma(y, z; x) := p_1(y_1) + f(y) + q_1(z_1) + g(z) + \langle x, \mathcal{A}^* y + \mathcal{B}^* z - c \rangle \\ + \frac{\sigma}{2} \|\mathcal{A}^* y + \mathcal{B}^* z - c\|^2, \end{aligned}$$

where $\sigma > 0$ is a penalty parameter, and x is the Lagrange multiplier. With initial point $(y^0, z^0, x^0) \in domp_1 \times domq_1 \times \mathcal{X}$, where $domp_1$ and $domq_1$ denote the domain of p_1 and q_1, the iterative scheme of sGS-ADMM algorithm for (A.8) in the $(k+1)$th $(k = 0, 1, 2, \ldots)$ iteration is

$$\left\{ \begin{array}{lcl} \tilde{y}_i^{k+1} & = & \arg\min\{\mathcal{L}_\sigma(y_{\leq i-1}^k, y_i, \tilde{y}_{\geq i+1}^{k+1}, z^k; x^k)\}, i = m, \cdots, 2, \\ y_i^{k+1} & = & \arg\min\{\mathcal{L}_\sigma(y_{\leq i-1}^{k+1}, y_i, \tilde{y}_{\geq i+1}^{k+1}, z^k; x^k)\}, i = 1, \cdots, m, \\ \tilde{z}_j^{k+1} & = & \arg\min\{\mathcal{L}_\sigma(y^{k+1}, z_{\leq j-1}^k, z_j, \tilde{z}_{\geq j+1}^{k+1}; x^k)\}, j = n, \cdots, 2, \\ z_j^{k+1} & = & \arg\min\{\mathcal{L}_\sigma(y^{k+1}, z_{\leq j-1}^{k+1}, z_j, \tilde{z}_{\geq j+1}^{k+1}; x^k)\}, j = 1, \cdots, n, \\ x^{k+1} & = & x^k - \tau\sigma(\mathcal{A}^* y^{k+1} + \mathcal{B}^* z^{k+1} - c), \end{array} \right.$$

where $y_{\leq i-1} := (y_1, \ldots, y_{i-1})$, $\tilde{y}_{\geq i+1} := (\tilde{y}_{i+1}, \ldots, \tilde{y}_m)$, $z_{\leq j-1} := (z_1, \ldots, z_{j-1})$, $\tilde{z}_{\geq j+1} := (\tilde{z}_{j+1}, \ldots, \tilde{z}_n)$ and τ is the step length. Now we present the convergence theorem of sGS-ADMM algorithm.

Theorem A.3. *Suppose that the solution set $\bar{W}$ to the KKT system of problem ([A.8](#)) is nonempty and the sequence $\{y^k, z^k, x^k\}$ is generated by the sGS-ADMM in the kth iteration. Let Σ_f, Σ_g, S and T be self-adjoint positive semidefinite linear operators such that $\Sigma_f + S + \sigma\mathcal{A}\mathcal{A}^* \succ 0$ and $\Sigma_g + T + \sigma\mathcal{B}\mathcal{B}^* \succ 0$. Then the sequence $\{y^k, z^k, x^k\}$ converges to a point in $\bar{W}$.*

A.3. Iterative scheme sGS-ADMM algorithm for DFMR

The each subproblems in Table 1 have closed-form solutions, which are obtained from the derivative of the augmented Lagrangian function or the properties of proximal mapping. Now let us look at the resulting subproblems in Table 1 one by one. The $u-$subproblem can be written as

$$
\begin{aligned}
u^{k+\frac{1}{2}} = \arg\min_{u} \Big\{ & \frac{1}{2}\|u\|_2^2 - u^{\mathrm{T}}y - \langle x_1^k, \mathbb{X}^{\mathrm{T}}u\rangle - \langle x_2^k, \mathbb{Z}^{\mathrm{T}}u\rangle + \frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}}u - w^k\|^2 \\
& + \frac{\sigma}{2}\|\mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v^k - \mathrm{vec}(D^k)\|^2 \Big\}.
\end{aligned}
$$

It is a quadratic form of u and has a unique closed-form solution

$$
u^{k+\frac{1}{2}} = (I + \sigma\mathbb{X}\mathbb{X}^{\mathrm{T}} + \sigma\mathbb{Z}\mathbb{Z}^{\mathrm{T}})^{-1}(y + \mathbb{X}x_1^k + \mathbb{Z}x_2^k - \sigma(\mathbb{X}C^{\mathrm{T}}v^k - \mathbb{X}\mathrm{vec}(D^k) - \mathbb{Z}w^k)).
$$

Similarly, the unique closed-form solution of the $v-$subproblem is

$$
v^{k+\frac{1}{2}} = \frac{1}{\sigma}(I + CC^{\mathrm{T}})^{-1}(Cx_1^k + x_3^k - \sigma(C\mathbb{X}^{\mathrm{T}}u^{k+\frac{1}{2}} - C\mathrm{vec}(D^k) + t^k)).
$$

The $D-$subproblem can be written as

$$
\begin{aligned}
D^{k+1} &= \arg\min_{D} \left\{ \delta(D; B_{\|\cdot\|_2(0;\lambda_1)}) + \frac{\sigma}{2}\|\mathbb{X}^{\mathrm{T}}u^{k+\frac{1}{2}} + C^{\mathrm{T}}v^{k+\frac{1}{2}} - \mathrm{vec}(D) - \frac{x_1^k}{\sigma}\|_2^2 \right\} \\
&= \arg\min_{D} \left\{ \delta(D; B_{\|\cdot\|_2(0;\lambda_1)}) + \frac{\sigma}{2}\|D - E\|_F^2 \right\},
\end{aligned}
$$

where $vec(E) = \mathbb{X}^{\mathrm{T}}u^{k+\frac{1}{2}} + C^{\mathrm{T}}v^{k+\frac{1}{2}} - x_1^k/\sigma$. By (A.7), the closed-form solution is $D^{k+1} = UDiag(e^*)V^{\mathrm{T}}$, where U, V, e satisfy the singular value decomposition of E, i.e., $E = U\Sigma V^{\mathrm{T}}$. The vector e contains the diagonal element of Σ, and $e^* = \Pi(e; B_{\|\cdot\|_\infty(0;\lambda_1)})$.

The $t-$subproblem can be written as

$$
\begin{aligned}
t^{k+1} &= \arg\min_{t} \left\{ \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) - \langle x_3^k, v^{k+\frac{1}{2}} + t\rangle + \frac{\sigma}{2}\|v^{k+\frac{1}{2}} + t\|_2^2 \right\} \\
&= \arg\min_{t} \left\{ \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) + \frac{\sigma}{2}\|v^{k+\frac{1}{2}} + t - x_3^k/\sigma\|_2^2 \right\}.
\end{aligned}
$$

By (A.3), the closed-form solution can be obtained from the soft-thresholding operator

$$
t^{k+1} = \Pi(x_3^k/\sigma - v^{k+\frac{1}{2}}; B_{\|\cdot\|_\infty(0;\lambda_2)})
$$

$$= (x_3^k/\sigma - v^{k+\frac{1}{2}}) - sign(x_3^k/\sigma - v^{k+\frac{1}{2}}) \cdot \max\left\{|v^{k+\frac{1}{2}} - x_3^k/\sigma| - \lambda_2, 0\right\}.$$

The $w-$subproblem can be written as

$$\begin{aligned} w^{k+1} &= \arg\min_w \left\{P^*(w) - \langle x_2^k, \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - w\rangle + \frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - w\|_2^2\right\} \\ &= \arg\min_w \left\{P^*(w) + \frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - w - x_2^k/\sigma\|_2^2\right\}. \end{aligned}$$

Applying the Moreau identity (A.2), we have

$$w^{k+1} = Prox_{P^*/\sigma}(\mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - \frac{x_2^k}{\sigma}) = (\mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - \frac{x_2^k}{\sigma}) - \frac{1}{\sigma} Prox_{\sigma P}(\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k).$$

The closed-form solution for $Prox_{\sigma P}(\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k)$ is given by

$$\begin{aligned} Prox_{\sigma P}(\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k) =& sign(x_{\sigma\lambda_4}(\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k)) \\ & \cdot \max\left\{|x_{\sigma\lambda_4}(\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k)| - \sigma\lambda_3, 0\right\}, \end{aligned}$$

where $x_{\sigma\lambda_4}(\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k) = \arg\min_x \{\sigma\lambda_4 \|A_p x\|_1 + \frac{1}{2}\|x - (\sigma \mathbb{Z}^{\mathrm{T}} u^{k+\frac{1}{2}} - x_2^k)\|_2^2\}$.

Finally, the stopping criterion *eta* for DFMR estimator is derived from the KKT condition.

$$\begin{aligned} \eta_P &= \max\left\{\frac{\|u^{k+1} - y - \mathbb{X}x_1^{k+1} - \mathbb{Z}x_2^{k+1}\|}{1 + \|u^{k+1}\| + \|x_2^{k+1}\|}, \frac{\|Cx_1^{k+1} + x_3^{k+1}\|}{1 + \|x_1^{k+1}\| + \|x_3^{k+1}\|}\right\}, \\ \eta_D &= \max\left\{\frac{\|v^{k+1} + t^{k+1}\|}{1 + \|v^{k+1}\| + \|t^{k+1}\|}, \frac{\|\mathbb{X}^{\mathrm{T}} u^{k+1} + C^{\mathrm{T}} v^{k+1} - \mathrm{vec}(D^{k+1})\|}{1 + \|\mathrm{vec}(D^{k+1})\|}, \right. \\ &\qquad \left. \frac{\|\mathbb{Z}^{\mathrm{T}} u^{k+1} - w^{k+1}\|}{1 + \|w^{k+1}\|}\right\}, \\ eta &= \max\{\eta_P, \eta_D\} < tol. \end{aligned}$$

The maximum number of iterations k is set to be 5000.

A.4. Iterative scheme of sGS-ADMM algorithm for DFMLR

The iterative scheme of sGS-ADMM algorithm for solving (2.8) is summarized in Table 12.

The D, w, t subproblems have the same solutions as in the DFMR. Now we give the closed-form solutions for $u-$subproblem and $v-$subproblem. The solution of $u-$subproblem is

$$\begin{aligned} u^{k+\frac{1}{2}} =& \frac{1}{\sigma}(I + \mathbb{X}\mathbb{X}^{\mathrm{T}} + \mathbb{Z}\mathbb{Z}^{\mathrm{T}})^{-1}\left(\mathbb{X}x_1^k + \mathbb{Z}x_2^k + x_3^k\right. \\ & \left. - \sigma\big(\mathbb{X}C^{\mathrm{T}} v^k - \mathbb{X}\mathrm{vec}(D^k) - \mathbb{Z}w^k\big) + s^k - y\right). \end{aligned}$$

Table 12

Iterative scheme of sGS-ADMM algorithm for solving (2.8)

```
Algorithm 2:
    Input: $X, Z, y$ and tolerance level $tol$. Choose $\lambda_1 > 0, \lambda_2 > 0, \lambda_3 > 0, \lambda_4 > 0$ and $\sigma > 0$.
    Let $\tau \in (0, (1+\sqrt{5})/2)$ be the step-length. Set the initial point $(u^0, v^0, \alpha^0, x^0)$.
    For    $k = 0, 1, \cdots$ , perform the following steps:
    Step 1a. (Backward GS sweep)   Compute $u^{k+\frac{1}{2}}$ and $v^{k+\frac{1}{2}}$,
        $u^{k+\frac{1}{2}} = \arg\min_{u} \mathcal{L}_\sigma(u, v^k, \alpha^k; x^k)$,
        $v^{k+\frac{1}{2}} = \arg\min_{v} \mathcal{L}_\sigma(u^{k+\frac{1}{2}}, v, \alpha^k; x^k)$.
    Step 1b. (Forward GS sweep)   Compute $u^{k+1}$ , $v^{k+1}$ and $\alpha^{k+1}$,
        $\alpha^{k+1} = \arg\min_{\alpha} \mathcal{L}_\sigma(u^{k+\frac{1}{2}}, v^{k+\frac{1}{2}}, \alpha; x^k)$,
        $v^{k+1} = \arg\min_{v} \mathcal{L}_\sigma(u^{k+\frac{1}{2}}, v, \alpha^{k+1}; x^k)$,
        $u^{k+1} = \arg\min_{u} \mathcal{L}_\sigma(u, v^{k+1}, \alpha^{k+1}; x^k)$.
    Step 2.   Update Lagrange multipliers $x_1^{k+1}, x_2^{k+1}, x_3^{k+1}$ and $x_4^{k+1}$,
        $x_1^{k+1} = x_1^k - \tau\sigma(\mathbb{X}^{\mathrm{T}} u^{k+1} + C^{\mathrm{T}} v^{k+1} - \mathrm{vec}(D^{k+1}))$,
        $x_2^{k+1} = x_2^k - \tau\sigma(\mathbb{Z}^{\mathrm{T}} u^{k+1} - w^{k+1})$,
        $x_3^{k+1} = x_3^k - \tau\sigma(u^{k+1} + s^{k+1} - y)$,
        $x_4^{k+1} = x_4^k - \tau\sigma(v^{k+1} + t^{k+1})$.
      If   $eta < tol$   stop
```

Similar to the $u-$subproblem, $v-$subproblem has a unique closed-form solution

$$v^{k+\frac{1}{2}} = \frac{1}{\sigma}(I + CC^{\mathrm{T}})^{-1}(Cx_1^k + x_4^k - \sigma(C\mathbb{X}^{\mathrm{T}} u^{k+\frac{1}{2}} - C\mathrm{vec}(D^k) + t^k)).$$

The stopping criterion *eta* for DFMLR estimator is also derived from the KKT condition.

$$\begin{aligned}
\eta_P &= \max\left\{ \frac{\|\mathbb{X}x_1^{k+1} + \mathbb{Z}x_2^{k+1} + x_3^{k+1}\|}{1 + \|x_1^{k+1}\| + \|x_2^{k+1}\| + \|x_3^{k+1}\|}, \frac{\|Cx_1^{k+1} + x_4^{k+1}\|}{1 + \|x_1^{k+1}\| + \|x_4^{k+1}\|} \right\}, \\
\eta_D &= \max\left\{ \frac{\|v^{k+1} + t^{k+1}\|}{1 + \|v^{k+1}\| + \|t^{k+1}\|}, \frac{\|\mathbb{X}^{\mathrm{T}} u^{k+1} + C^{\mathrm{T}} v^{k+1} - \mathrm{vec}(D^{k+1})\|}{1 + \|\mathrm{vec}(D^{k+1})\|}, \right. \\
&\qquad \left. \frac{\|\mathbb{Z}^{\mathrm{T}} u^{k+1} - w^{k+1}\|}{1 + \|w^{k+1}\|}, \frac{\|u^{k+1} + s^{k+1} - y\|}{1 + \|u^{k+1}\| + \|s^{k+1}\|} \right\}, \\
eta &= \max\{\eta_P, \eta_D\} < tol.
\end{aligned}$$

The maximum number of iterations k was set to be 5000.

A.5. Proofs

Proof of Theorem 2.1. The global convergence of sGS-ADMM algorithm is established by [5]. For the problem (2.5) in our paper is a convex optimization problem which has better structures where $I + \sigma\mathbb{X}\mathbb{X}^{\mathrm{T}} + \sigma\mathbb{Z}\mathbb{Z}^{\mathrm{T}}$ and $I + CC^{\mathrm{T}}$ are positive definite matrices. The global convergence of the sGS-ADMM algorithm for solving problem (2.5) is easily satisfied. □

In order to prove the Theorem 2.2, we established two lemmas which give the upper bound of the KKT system of the iterative point and investigate the distance between iterative points and the optimal solution, respectively. We give the definition of metric subregularity from [7]. Let $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ be a multi-valued mapping. Denote its inverse by $\mathcal{F}^{-1}$. Define the graph of multi-valued functions $\mathcal{F}$ as follows

$$graph\mathcal{F} := \{(x, y) \in \mathcal{X} \times \mathcal{Y} | y \in \mathcal{F}(x)\}.$$

Definition A.4. *A multi-valued mapping $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ is said to be metric subregular at $\bar{x} \in \mathcal{X}$ for $\bar{y} \in \mathcal{Y}$ with modulus $\kappa > 0$ if $(\bar{x}, \bar{y}) \in gragh\mathcal{F}$ and there exist neighborhood $\mathcal{U}$ of $\bar{x}$ and $\mathcal{V}$ of $\bar{y}$ such that*

$$dist(x, \mathcal{F}^{-1}(\bar{y})) \le \kappa dist(\bar{y}, \mathcal{F}(x) \cap \mathcal{V}), \forall x \in \mathcal{U}.$$

Let $\Theta := \mathbb{R}^n \times \mathbb{R}^{(m-1)q} \times \mathcal{Y} \times \mathcal{X}$, and define the KKT mapping $R : \Theta \to \Theta$ as

$$R(\theta) := \begin{pmatrix} u - y - \mathbb{X}x_1 - \mathbb{Z}x_2 \\ -Cx_1 - x_3 \\ w - Prox_{P^*}(w - x_2) \\ D - Prox_{\delta_{B_{\|\cdot\|_2 \le \lambda_1}}}(D - \Xi) \\ t - Prox_{\delta_{B_{\|\cdot\|_\infty \le \lambda_2}}}(t + x_3) \\ \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D) \\ \mathbb{Z}^{\mathrm{T}}u - w \\ v + t \end{pmatrix}, \quad \forall \theta \in \Theta,$$

where $vec(\Xi) = x_1$. We know that $R(\theta) = 0$ is equal to $\theta \in \bar{\Omega}$. According to Moreau identity (A.2) $Prox_{\delta_{B_{\|\cdot\|_2 \le \lambda_1}}}(D - \Xi) + Prox_{\lambda_1\|\cdot\|_*}(D - \Xi) = D - \Xi$, we have $D - Prox_{\delta_{B_{\|\cdot\|_2 \le \lambda_1}}}(D - \Xi) = \Xi + Prox_{\lambda_1\|\cdot\|_*}(D - \Xi)$. Thus we obtain that

$$R(\theta) := \begin{pmatrix} u - y - \mathbb{X}x_1 - \mathbb{Z}x_2 \\ -Cx_1 - x_3 \\ w - Prox_{P^*}(w - x_2) \\ D - Prox_{\delta_{B_{\|\cdot\|_2 \le \lambda_1}}}(D - \Xi) \\ t - Prox_{\delta_{B_{\|\cdot\|_\infty \le \lambda_2}}}(t + x_3) \\ \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D) \\ \mathbb{Z}^{\mathrm{T}}u - w \\ v + t \end{pmatrix} = \begin{pmatrix} u - y - \mathbb{X}x_1 - \mathbb{Z}x_2 \\ -Cx_1 - x_3 \\ w - Prox_{P^*}(w - x_2) \\ \Xi + Prox_{\lambda_1\|\cdot\|_*}(D - \Xi) \\ t - Prox_{\delta_{B_{\|\cdot\|_\infty \le \lambda_2}}}(t + x_3) \\ \mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v - \mathrm{vec}(D) \\ \mathbb{Z}^{\mathrm{T}}u - w \\ v + t \end{pmatrix}.$$

Define $\kappa_1 := (12\sigma + 4\tau)\lambda_{\max}(\mathbb{X}^{\mathrm{T}}\mathbb{X}), \kappa_2 := (8\sigma + 4\tau)\lambda_{\max}(\mathbb{Z}^{\mathrm{T}}\mathbb{Z})$, $\kappa_3 := (12\sigma + 4\tau)\lambda_{\max}(C^{\mathrm{T}}C)$ and $\kappa_4 := \max\{\kappa_1 + \kappa_3 + 2\sigma + 2\tau + 1/\sigma, \kappa_2 + 2\sigma + 2\tau + 1/\sigma, 14\sigma + 5\tau + 1/\sigma\}$. Let $\mathcal{H}_0$ be the block-diagonal linear operator defined by $\mathcal{H}_0 := \kappa_4 Diag(0, CC^{\mathrm{T}}, \Sigma_I, (\tau^2\sigma)^{-1}I_x)$. We provide the following lemma which is useful in proving Theorem 2.2.

Lemma A.5. *Let $\{\theta^k := (u^k, v^k, \alpha^k, x^k)\}$ is generated by the sGS-ADMM. Then for any $k \ge 1$,*

$$\|\theta^{k+1} - \theta^k\|_{\mathcal{H}_0}^2 \ge \|R(\theta^{k+1})\|^2. \tag{A.9}$$

Proof. The optimal condition for every subproblem in Table 1 is

$$\begin{cases} 0 & = & u^{k+1} - y - \mathbb{X}x_1^k - \mathbb{Z}x_2^k + \sigma\mathbb{X}(\mathbb{X}^{\mathrm{T}}u^{k+1} + C^{\mathrm{T}}v^k - \mathrm{vec}(D^k)) \\ & & +\sigma\mathbb{Z}(\mathbb{Z}^{\mathrm{T}}u^{k+1} - w^k), \\ 0 & = & -Cx_1^k - x_3^k + \sigma C(\mathbb{X}u^{k+1} + C^{\mathrm{T}}v^{k+1} - \mathrm{vec}(D^k)) + \sigma(v^{k+1} + t^k), \\ 0 & \in & \partial P^*(w^{k+1}) + x_2^k - \sigma(\mathbb{Z}^{\mathrm{T}}u^{k+1} - w^{k+1}), \\ 0 & \in & \partial\delta_{B_{\|\cdot\|_2 \le \lambda_1}}(D^{k+1}) + \Xi^k - \sigma\Lambda^{k+1}, \\ 0 & \in & \partial\delta_{B_{\|\cdot\|_\infty \le \lambda_2}}(t^{k+1}) - x_3^k + \sigma(v^{k+1} + t^{k+1}), \end{cases}$$

where $vec(\Lambda^{k+1}) = \mathbb{X}u^{k+1} + C^{\mathrm{T}}v^{k+1} - \mathrm{vec}(D^{k+1})$. We obtain from the definition of $R(\cdot)$ that

$$\begin{aligned} \|R(\theta^{k+1})\|^2 \le & \|u^{k+1} - y - \mathbb{X}x_1^{k+1} - \mathbb{Z}x_2^{k+1}\|^2 + \|Cx_1^{k+1} + x_3^{k+1}\|^2 \\ & + \|\sigma(\mathbb{Z}^{\mathrm{T}}u^{k+1} - w^{k+1}) - x_2^k + x_2^{k+1}\|^2 + \|\sigma\Lambda^{k+1} - \Xi^k + \Xi^{k+1}\|_F^2 \\ & + \|\sigma(v^{k+1} + t^{k+1}) - x_3^k + x_3^{k+1}\|^2 + \|v^{k+1} + t^{k+1}\|^2 \\ & + \|\mathbb{X}^{\mathrm{T}}u^{k+1} + C^{\mathrm{T}}v^{k+1} - \mathrm{vec}(D^{k+1})\|^2 + \|\mathbb{Z}^{\mathrm{T}}u^{k+1} - w^{k+1}\|^2. \end{aligned}$$

By schemes of Lagrange multipliers and the definition of Λ^{k+1}, we have

$$\begin{aligned} & \|R(\theta^{k+1})\|^2 \\ \le & 12\sigma^2\lambda_{max}(\mathbb{X}^{\mathrm{T}}\mathbb{X})\|v^k - v^{k+1}\|_{CC^{\mathrm{T}}}^2 + 8\sigma^2\lambda_{max}(\mathbb{Z}^{\mathrm{T}}\mathbb{Z})\|w^{k+1} - w^k\|^2 \\ & + (12\sigma^2\lambda_{max}(\mathbb{X}^{\mathrm{T}}\mathbb{X}) + 12\sigma^2\lambda_{max}(C^TC))\|\mathrm{vec}(D^{k+1}) - \mathrm{vec}(D^k)\|^2 \\ & + 12\sigma^2\|t^k - t^{k+1}\|^2 + ((12\sigma^2 + 4\tau\sigma)\lambda_{max}(\mathbb{X}^{\mathrm{T}}\mathbb{X}) \\ & + (12\sigma^2 + 3\tau\sigma)\lambda_{max}(C^TC) + 2\sigma^2 + 2\tau\sigma + 1)\|(\tau\sigma)^{-1}(x_1^k - x_1^{k+1})\|^2 \\ & + ((8\sigma^2 + 4\tau\sigma)\lambda_{max}(\mathbb{Z}^{\mathrm{T}}\mathbb{Z}) + 2\sigma^2 + 2\tau\sigma + 1)\|(\tau\sigma)^{-1}(x_2^k - x_2^{k+1})\|^2 \\ & + (14\sigma^2 + 5\tau\sigma + 1)\|(\tau\sigma)^{-1}(x_3^k - x_3^{k+1})\|^2 \\ \le & \kappa_1\|v^k - v^{k+1}\|_{CC^{\mathrm{T}}}^2 + \kappa_2\|w^{k+1} - w^k\|^2 + 12\sigma^2\|t^k - t^{k+1}\|^2 \\ & + (\kappa_1 + \kappa_3 + 2\sigma^2 + 2\tau\sigma + 1)\|(\tau\sigma)^{-1}(x_1^k - x_1^{k+1})\|^2 \\ & + (\kappa_2 + 2\sigma^2 + 2\tau\sigma + 1)\|(\tau\sigma)^{-1}(x_2^k - x_2^{k+1})\|^2 \\ & + (14\sigma^2 + 5\tau\sigma + 1)\|(\tau\sigma)^{-1}(x_3^k - x_3^{k+1})\|^2 \\ & + (\kappa_1 + \kappa_3)\|\mathrm{vec}(D^{k+1}) - \mathrm{vec}(D^k)\|^2. \end{aligned}$$

Thus we can immediately imply (A.9). □

Define $t_\tau := \frac{1}{2}(1 - \tau + \min\{\tau, \tau^{-1}\})$ and the self-adjoint linear operator $\mathcal{H} := Diag(\frac{1}{2}I, 2t_\tau\tau\sigma(I + CC^{\mathrm{T}}), 2t_\tau\tau\sigma\Sigma_I + \frac{1}{2}\Sigma_\alpha, t_\tau(\tau^2\sigma)^{-1}I_x) + \frac{1}{4}t_\tau\sigma\Phi\Phi^*$. We easily know that $1/4 \le s_\tau \le 5/4$ and $0 \le t_\tau \le 1/2$. We next give inequality which plays a key role in deriving the Q-linear rate of convergence for the sGS-ADMM algorithm of DFMR estimator.

Lemma A.6. *Let $\{\theta^k := (u^k, v^k, \alpha^k, x^k)\}$ be an infinite sequence generated by the sGS-ADMM. Then for any $\bar{\theta} = (\bar{u}, \bar{v}, \bar{\alpha}, \bar{x}) \in \bar{\Omega}$ and any $k \ge 1$,*

$$\|\theta^{k+1} - \bar{\theta}\|_{\mathcal{M}}^2 \le \|\theta^k - \bar{\theta}\|_{\mathcal{M}}^2 - \|\theta^{k+1} - \theta^k\|_{\mathcal{H}}^2. \qquad \text{(A.10)}$$

Consequently, we have

$$dist^2_{\mathcal{M}}(\theta^{k+1},\bar{\Omega}) \le dist^2_{\mathcal{M}}(\theta^{k},\bar{\Omega}) - \|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}}. \tag{A.11}$$

Proof. For any θ and θ', we define the function $J(\theta,\theta') := (\tau\sigma)^{-1}\|x-x'\|^2 + \sigma\|v-v'\|^2_{I+CC^{\mathrm{T}}} + \sigma\|\Sigma_I(\alpha-\alpha')\|^2$. Let $\bar{\theta}=(\bar{u},\bar{v},\bar{\alpha},\bar{x})\in\bar{\Omega}$. Following Appendix B of [14], for any $k\ge 1$ the inequality holds that

$$\begin{aligned}
&J(\theta^{k+1},\bar{\theta}) + (1-\min\{\tau,\tau^{-1}\})\sigma\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2\\
&-[J(\theta^{k},\bar{\theta}) + (1-\min\{\tau,\tau^{-1}\})\sigma\|\Phi^*(u^{k},v^{k},\alpha^{k},0)\|^2]\\
\le& -\tau(1-\tau+\min\{\tau,\tau^{-1}\})\sigma(\|v^{k+1}-v^k\|^2_{I+CC^{\mathrm{T}}} + \|\Sigma_I(\alpha^{k+1}-\alpha^k)\|^2)\\
&-(1-\tau+\min\{\tau,\tau^{-1}\})\sigma\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2\\
&-2\|u^{k+1}-\bar{u}\|^2 - 2\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}.
\end{aligned} \tag{A.12}$$

By reorganizing the terms in (A.12), we obtain

$$\begin{aligned}
&(\tau\sigma)^{-1}\|x^{k+1}-\bar{x}\|^2 + \sigma\|v^{k+1}-\bar{v}\|^2_{I+CC^{\mathrm{T}}} + \sigma\|\Sigma_I(\alpha^{k+1}-\bar{\alpha})\|^2\\
&+s_\tau\sigma\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2 + \|u^{k+1}-\bar{u}\|^2 + \|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}\\
\le&(\tau\sigma)^{-1}\|x^{k}-\bar{x}\|^2 + \sigma\|v^{k}-\bar{v}\|^2_{I+CC^{\mathrm{T}}} + \sigma\|\Sigma_I(\alpha^{k}-\bar{\alpha})\|^2\\
&+s_\tau\sigma\|\Phi^*(u^{k},v^{k},\alpha^{k},0)\|^2 + \|u^{k}-\bar{u}\|^2 + \|\alpha^{k}-\bar{\alpha}\|^2_{\Sigma_\alpha}\\
&-\{2t_\tau\tau\sigma[\|v^{k+1}-v^k\|^2_{I+CC^{\mathrm{T}}} + \|\Sigma_I(\alpha^{k+1}-\alpha^k)\|^2] + \|u^{k+1}-\bar{u}\|^2\\
&+\|u^k-\bar{u}\|^2 + \|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha} + \|\alpha^k-\bar{\alpha}\|^2_{\Sigma_\alpha} + t_\tau\sigma\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2\\
&+\frac{1}{2}t_\tau\sigma[\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2 + \|\Phi^*(u^{k},v^{k},\alpha^{k},0)\|^2]\}.
\end{aligned}$$

Using equalities

$$\begin{aligned}
\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0) &= (\tau\sigma)^{-1}(x^k-x^{k+1}),\\
\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2 &= \|\theta^{k+1}-\bar{\theta}\|^2_{\Phi\Phi^*}
\end{aligned}$$

and inequalities

$$\begin{aligned}
&\|u^{k+1}-\bar{u}\|^2 + \|u^k-\bar{u}\|^2 \ge \frac{1}{2}\|u^{k+1}-u^k\|^2,\\
&\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha} + \|\alpha^k-\bar{\alpha}\|^2_{\Sigma_\alpha} \ge \frac{1}{2}\|\alpha^{k+1}-\alpha^k\|^2_{\Sigma_\alpha},\\
&\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2 + \|\Phi^*(u^{k},v^{k},\alpha^{k},0)\|^2 \ge \frac{1}{2}\|\theta^{k+1}-\theta^k\|^2_{\Phi\Phi^*},
\end{aligned}$$

we obtain

$$\begin{aligned}
&(\tau\sigma)^{-1}\|x^{k+1}-\bar{x}\|^2 + \sigma\|v^{k+1}-\bar{v}\|^2_{I+CC^{\mathrm{T}}} + \sigma\|\alpha^{k+1}-\bar{\alpha}\|^2\\
&+s_\tau\sigma\|\theta^{k+1}-\bar{\theta}\|^2_{\Phi\Phi^*} + \|u^{k+1}-\bar{u}\|^2 + \|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}\\
\le&(\tau\sigma)^{-1}\|x^{k}-\bar{x}\|^2 + \sigma\|v^{k}-\bar{v}\|^2_{I+CC^{\mathrm{T}}} + \sigma\|\alpha^{k}-\bar{\alpha}\|^2 + s_\tau\sigma\|\theta^{k}-\bar{\theta}\|^2_{\Phi\Phi^*}
\end{aligned}$$

$$
\begin{aligned}
&+\|u^k-\bar{u}\|^2+\|\alpha^k-\bar{\alpha}\|^2_{\Sigma_\alpha}-\{2t_\tau\tau\sigma[\|v^{k+1}-v^k\|^2_{I+CC^{\mathrm{T}}}\\
&+\|\alpha^{k+1}-\alpha^k\|^2]+\frac{1}{2}\|u^{k+1}-u^k\|^2+\frac{1}{2}\|\alpha^{k+1}-\alpha^k\|^2_{\Sigma_\alpha}\\
&+t_\tau(\tau^2\sigma)^{-1}\|x^k-x^{k+1}\|^2+\frac{1}{4}t_\tau\sigma\|\theta^{k+1}-\theta^k\|^2_{\Phi\Phi^*}\}.
\end{aligned}
$$

It shows that (A.10) holds. Note that $\bar{\Omega}$ is a nonempty closed convex set and (A.10) holds for any $\bar{\theta}\in\bar{\Omega}$, we immediately get (A.11). □

Based on Lemma A.5 and Lemma A.6, we give a specific proof for the Q-linear rate of convergence of the sGS-ADMM algorithm.

Proof of Theorem 2.2. We know that L_1 norm and fused Lasso regularization are polyhedral convex functions [27, 40], their Fenchel conjugate functions are also polyhedral convex functions. According to [20], $Prox_{P^*}(\cdot)$ and $Prox_{\delta_{B_{\|\cdot\|_\infty\le\lambda_2}}}(\cdot)$ are piecewise polyhedral. The multi-valued mapping $\partial\|\cdot\|_*:\mathbb{R}^{m\times q}\to\mathbb{R}^{m\times q}$ is metrically subregular at the KKT point for origin [51]. Thus the multi-valued mapping $R(\theta)$ is metrically subregular at the KKT point for origin. There exist positive constants $\hat{\eta}>0$ and $\hat{\delta}>0$ such that $dist(\theta^k,\bar{\Omega})\le\hat{\eta}\|R(\theta^k)\|,\forall\theta\in\{\theta:\|\theta-\bar{\theta}\|\le\hat{\delta}\}$. According to Lemma A.5, it holds that $\|R(\theta^{k+1})\|^2\le\|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}_0}$. Thus, we get that for all $k\ge1$, $dist^2(\theta^{k+1},\bar{\Omega})\le\hat{\eta}^2\|R(\theta^{k+1})\|^2\le\hat{\eta}^2\|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}_0}$. We have for all $k\ge1$, $\|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}}\ge0$ and

$$
\begin{aligned}
\|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}}&\ge\min\{2\tau,1\}t_\tau\kappa_4^{-1}\|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}_0}\\
&\ge\min\{2\tau,1\}t_\tau\kappa_4^{-1}\hat{\eta}^{-2}dist^2(\theta^{k+1},\bar{\Omega})\\
&\ge\kappa dist^2_{\mathcal{M}}(\theta^{k+1},\bar{\Omega}),
\end{aligned}
\tag{A.13}
$$

where $\kappa=\min\{2\tau,1\}t_\tau\kappa_4^{-1}\hat{\eta}^{-2}>0$. According to (A.11) and (A.13), we have $dist^2_{\mathcal{M}}(\theta^{k+1},\bar{\Omega})-dist^2_{\mathcal{M}}(\theta^k,\bar{\Omega})\le-\|\theta^{k+1}-\theta^k\|^2_{\mathcal{H}}\le-\kappa dist^2_{\mathcal{M}}(\theta^{k+1},\bar{\Omega})$, it holds that $(1+\kappa)dist^2_{\mathcal{M}}(\theta^{k+1},\bar{\Omega})\le dist^2_{\mathcal{M}}(\theta^k,\bar{\Omega})$. Denote $\mu=(1+\kappa)^{-1}<1$, The proof of Theorem 2.2 has been completed. □

The proof of Theorem 2.3 is similar to Theorem 2.2, we will omit it.

Proof of Theorem 3.1. In our proof, we will divide the proof into two steps: (1) prove that $||\hat{B}-B^*||^2_F+||\hat{\gamma}-\gamma^*||^2_2$ has a upper bound; (2) prove that $||\hat{B}-B^*||^2_F+||\hat{\gamma}-\gamma^*||^2_2$ converges in probability to zero.

Step 1. According to the definition of $\hat{B},\hat{\gamma}$, we obtain that

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^n(y_i-\langle X_i,\hat{B}\rangle-\langle z_i,\hat{\gamma}\rangle)^2+\lambda_1||\hat{B}||_*+\lambda_2||Cvec(\hat{B})||_1+\lambda_3||\hat{\gamma}||_1\\
&+\lambda_4||A_p\hat{\gamma}||_1\\
\le&\frac{1}{2}\sum_{i=1}^n\varepsilon_i^2+\lambda_1||B^*||_*+\lambda_2||Cvec(B^*)||_1+\lambda_3||\gamma^*||_1+\lambda_4||A_p\gamma^*||_1,
\end{aligned}
$$

which yields that

$$
\frac{1}{2}\sum_{i=1}^n\Big(\varepsilon_i-\langle X_i,\hat{B}-B^*\rangle-\langle z_i,\hat{\gamma}-\gamma^*\rangle\Big)^2
$$

$$\leq \frac{1}{2}\sum_{i=1}^{n}\varepsilon_i^2 + \lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1. \tag{A.14}$$

By the formulation $(a+b)^2 = a^2 + b^2 + 2ab$, we have

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}\Big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\Big)^2 + \sum_{i=1}^{n}\varepsilon_i\Big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\Big)\\ \leq&\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1.\end{aligned} \tag{A.15}$$

Therefore, we obtain

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}\big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\big)^2\\ =&\frac{1}{2}\sum_{i=1}^{n}\big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle - \varepsilon_i + \varepsilon_i\big)^2\\ \leq&\sum_{i=1}^{n}\big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle - \varepsilon_i\big)^2 + \sum_{i=1}^{n}\varepsilon_i^2\\ \leq&2\sum_{i=1}^{n}\varepsilon_i^2 + 2\lambda_1||B^*||_* + 2\lambda_2||Cvec(B^*)||_1 + 2\lambda_3||\gamma^*||_1 + 2\lambda_4||A_p\gamma^*||_1.\end{aligned}$$

By Cauchy's inequality and Condition 4, we have

$$\begin{aligned}&\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1\\ \leq&\lambda_1(r^*\|B^*\|_2) + \lambda_2(||Cvec(B^*)||_\infty s_1^*) + \lambda_3(s_2^*\|\gamma^*||_\infty) + \lambda_4(s_3^*\|A_p\gamma^*||_\infty)\\ \leq&4\sqrt{n}.\end{aligned}$$

According to the above two inequalities, we infer that

$$\frac{1}{2}\sum_{i=1}^{n}\big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\big)^2 \leq 2\sum_{i=1}^{n}\varepsilon_i^2 + 8\sqrt{n}. \tag{A.16}$$

By Chebyshev's inequality and Condition 3, we obtain that

$$\begin{aligned}P\Big(\sum_{i=1}^{n}\varepsilon_i^2 \geq 2n\sigma_0^2/k^2\Big) &\leq e^{-2n\sigma_0^2/k^2}\mathbb{E}e^{\sum_{i=1}^{n}\varepsilon_i^2/k^2}\\ &= e^{-2n\sigma_0^2/k^2}\cdot e^{\sum_{i=1}^{n}\log\mathbb{E}(e^{\varepsilon_i^2/k^2})}\\ &\leq e^{-2n\sigma_0^2/k^2 + n\log(1+\sigma_0^2/k^2)}\\ &\leq e^{-2n\sigma_0^2/k^2 + n\sigma_0^2/k^2}\\ &= e^{-n\sigma_0^2/k^2}.\end{aligned} \tag{A.17}$$

Notice that

$$\begin{aligned}\langle X_i, \hat{B} - B^*\rangle\langle z_i, \hat{\gamma} - \gamma^*\rangle =&\text{trace}\Big(\big((\hat{\gamma} - \gamma^*)^T z_i\big) \otimes \big((\hat{B} - B^*)^T X_i\big)\Big)\\ =&\text{trace}\Big(\big((\hat{\gamma} - \gamma^*)^T \otimes (\hat{B} - B^*)^T\big)\big(z_i \otimes X_i\big)\Big)\\ =&\text{trace}\Big(\big((\hat{\gamma} - \gamma^*) \otimes (\hat{B} - B^*)\big)^T\big(z_i \otimes X_i\big)\Big)\\ =&\langle z_i \otimes X_i, (\hat{\gamma} - \gamma^*) \otimes (\hat{B} - B^*)\rangle\\ \leq&\|z_i \otimes X_i\|_F \|(\hat{\gamma} - \gamma^*) \otimes (\hat{B} - B^*)\|_F\\ \leq&\|z_i \otimes X_i\|_F \|\hat{\gamma} - \gamma^*\|_2 \|\hat{B} - B^*\|_F.\end{aligned}$$

From the above inequality, Conditions 1-2 and the condition (3.1), we conclude that

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^{n}\big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\big)^2\\ =&\frac{1}{2}\sum_{i=1}^{n}\langle X_i, \hat{B} - B^*\rangle^2 + \frac{1}{2}\sum_{i=1}^{n}\langle z_i, \hat{\gamma} - \gamma^*\rangle^2 + \sum_{i=1}^{n}\langle X_i, \hat{B} - B^*\rangle\langle z_i, \hat{\gamma} - \gamma^*\rangle\\ \geq&\frac{n}{2}\underline{C}_X||\hat{B} - B^*||_F^2 + \frac{n}{2}\underline{C}_Z||\hat{\gamma} - \gamma^*||^2\\ &- n||\frac{1}{n}\sum_{i=1}^{n}(z_i \otimes X_i)||_F \cdot ||\hat{B} - B^*||_F \cdot ||\hat{\gamma} - \gamma^*||\\ \geq&\frac{n}{2}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2)\\ &- \frac{n}{2}\min\{\underline{C}_X, \underline{C}_Z\} \cdot \frac{1}{2}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2)\\ \geq&\frac{n}{4}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2).\end{aligned}\tag{A.18}$$

Combing (A.16), (A.17) and (A.18) yields

$$||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2 \leq \frac{8}{\min\{\underline{C}_X, \underline{C}_Z\}} \cdot \Big(3/\sqrt{n} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\Big) \leq c^2 \tag{A.19}$$

with probability at least $1 - e^{-n\sigma_0^2/k^2}$.

Step 2. According to (A.19), we have already seen that $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2$ has a upper bound. Now we prove that $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2$ converges in probability to zero.

By (A.15) and and (A.18), we know

$$\begin{aligned}&\frac{1}{4}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2)\\ \leq&\frac{1}{n}|\sum_{i=1}^{n}\varepsilon_i\Big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\Big)|\end{aligned}$$

$$+\frac{\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1}{n}. \tag{A.20}$$

For simplicity, we denote $U = \hat{B} - B^*, v = \hat{\gamma} - \gamma^*$. We estimate the upper bound $\frac{1}{n}|\sum_{i=1}^n \varepsilon_i\Big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\Big)|$.

$$\begin{aligned}
&\sup_{\|U\|_F^2+\|v\|^2\le c} \frac{1}{n}|\sum_{i=1}^n \varepsilon_i\Big(\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle\Big)| \\
&= \sup_{\|U\|_F^2+\|v\|^2\le c} \frac{1}{n}|\sum_{i=1}^n \varepsilon_i\Big(\langle X_i, U\rangle + \langle z_i, v\rangle\Big)| \\
&\le \sup_{\|U\|_F^2+\|v\|^2\le c} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle X_i, U\rangle| + \sup_{\|U\|_F^2+\|v\|^2\le c} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle z_i, v\rangle|.
\end{aligned}$$

Condition 3 shows that the ε obeys the sub-Guassian distribution. Applying Hoeffding inequality for every $\tau \ge 0$ yields

$$P(|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle X_i, U\rangle| > \tau) \le e^{-\frac{n\tau^2}{8(k^2+\sigma_0^2)\overline{C}_X\|U\|_F^2}}.$$

By taking $\tau = \sqrt{\frac{(k^2+\sigma_0^2)\|U\|_F^2\overline{C}_X mq}{n}}$, we have

$$\mathbb{P}\Big(\sup_{\|U\|_F^2+\|v\|^2\le c} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle X_i, U\rangle| \ge \sqrt{\frac{(k^2+\sigma_0^2)\|U\|_F^2\overline{C}_X mq}{n}}\Big) \le c_1 e^{-c_2 mq} \tag{A.21}$$

Similarly,

$$\mathbb{P}\Big(\sup_{\|U\|_F^2+\|v\|^2\le c} |\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle z_i, v\rangle| \ge \sqrt{\frac{(k^2+\sigma_0^2)\|v\|_2^2\overline{C}_Z p}{n}}\Big) \le c_3 e^{-c_4 p}, \tag{A.22}$$

where $\{c_i\}_{i=1}^4$ are positive constants.

Combing (A.20), (A.21) and (A.22) leads to

$$\begin{aligned}
&\frac{1}{4}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2) \\
\le&\frac{1}{n}|\sum_{i=1}^n \varepsilon_i\langle X_i, \hat{B} - B^*\rangle + \langle z_i, \hat{\gamma} - \gamma^*\rangle| \\
&+\frac{\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1}{n} \\
\le& \sup_{\|U\|_F^2+\|v\|^2\le c^2} \frac{1}{n}|\sum_{i=1}^n \varepsilon_i\langle X_i, \hat{B} - B^*\rangle| + \sup_{\|U\|_F^2+\|v\|^2\le c^2} \frac{1}{n}|\sum_{i=1}^n \varepsilon_i\langle z_i, \hat{\gamma} - \gamma^*\rangle| \\
&+\frac{\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1}{n}
\end{aligned}$$

$$\begin{aligned}
\leq & \sqrt{\frac{(k^2+\sigma_0^2)\overline{C}_X mq}{n}}\|U\|_F + \sqrt{\frac{(k^2+\sigma_0^2)\overline{C}_X p}{n}}\|v\|_2 \\
& + \frac{\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1}{n}.
\end{aligned}$$

Due to

$$\begin{aligned}
\sqrt{\frac{(k^2+\sigma_0^2)\overline{C}_X mq}{n}}\|U\|_F &= 2\sqrt{2}\sqrt{\frac{(k^2+\sigma_0^2)\overline{C}_X mq}{n\min\{\underline{C}_X,\underline{C}_Z\}}} \cdot \frac{\sqrt{\min\{\underline{C}_X,\underline{C}_Z\}}}{2\sqrt{2}}\|U\|_F \\
&\leq \frac{2(k^2+\sigma_0^2)\overline{C}_X mq}{n\min\{\underline{C}_X,\underline{C}_Z\}} + \frac{1}{8}\min\{\underline{C}_X,\underline{C}_Z\}\|U\|_F^2
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{\frac{(k^2+\sigma_0^2)\overline{C}_Z p}{n}}\|v\|_2 &= 2\sqrt{2}\sqrt{\frac{(k^2+\sigma_0^2)\overline{C}_Z p}{n\min\{\underline{C}_X,\underline{C}_Z\}}} \cdot \frac{\sqrt{\min\{\underline{C}_X,\underline{C}_Z\}}}{2\sqrt{2}}\|v\|_2 \\
&\leq \frac{2(k^2+\sigma_0^2)\overline{C}_Z p}{n\min\{\underline{C}_X,\underline{C}_Z\}} + \frac{1}{8}\min\{\underline{C}_X,\underline{C}_Z\}\|v\|_2^2,
\end{aligned}$$

we obtain that

$$\begin{aligned}
||\hat{B}-B^*||_F^2 + ||\hat{\gamma}-\gamma^*||_2^2 \leq & \frac{16(k^2+\sigma_0^2)}{\min^2\{\underline{C}_X,\underline{C}_Z\}}\frac{\overline{C}_X mq + \overline{C}_Z p}{n} \\
+ & \frac{8(\lambda_1 r^*\|B^*\|_2 + \lambda_2||Cvec(B^*)||_\infty s_1^* + \lambda_3 s_2^*\|\gamma^*||_\infty + \lambda_4 s_3^*\|A_p\gamma^*||_\infty)}{n\min\{\underline{C}_X,\underline{C}_Z\}}
\end{aligned}$$

with probability at least $1-e^{-n\sigma_0^2/k^2} - c_1e^{-c_2mq} - c_3e^{-c_4p}$, where $\{c_i\}_{i=1}^4$ are positive constants. Then, we get the inequality in Theorem 3.1. As a direct consequence, it follows that $||\hat{B}-B^*||_F^2 + ||\hat{\gamma}-\gamma^*||_2^2$ converges in probability to zero under the Condition 4 and assumption. □

Proof of Theorem 3.2. For simplicity, denote

$$H(B,\gamma) = \sum_{i=1}^n \log(1+e^{\langle X_i,B\rangle + \langle z_i,\gamma\rangle}) - y_i(\langle X_i,B\rangle + \langle z_i,\gamma\rangle).$$

According to Conditions 5-6, we obtain

$$\begin{aligned}
& \sum_{i=1}^n K\langle X_i,\hat{B}-B^*\rangle^2 + (\hat{\gamma}-\gamma^*)^T\sum_{i=1}^n Kz_iz_i^T(\hat{\gamma}-\gamma^*) \\
& + 2\sum_{i=1}^n \langle X_i,\hat{B}-B^*\rangle\langle Kz_i,\hat{\gamma}-\gamma^*\rangle \\
\geq & n\underline{C}_X||\hat{B}-B^*||_F^2 + n\underline{C}_Z||\hat{\gamma}-\gamma^*||^2 \\
& - 2n||\frac{1}{n}\sum_{i=1}^n (Kz_i\otimes X_i)||_F \cdot ||\hat{B}-B^*||_F \cdot ||\hat{\gamma}-\gamma^*||
\end{aligned}$$

$$
\begin{aligned}
&\geq n\min\{\underline{C}_X,\underline{C}_Z\}(||\hat{B}-B^*||_F^2+||\hat{\gamma}-\gamma^*||^2)\\
&\quad -n\min\{\underline{C}_X,\underline{C}_Z\}\cdot\frac{1}{2}(||\hat{B}-B^*||_F^2+||\hat{\gamma}-\gamma^*||^2)\\
&\geq\frac{n}{2}\min\{\underline{C}_X,\underline{C}_Z\}(||\hat{B}-B^*||_F^2+||\hat{\gamma}-\gamma^*||^2). \qquad (A.23)
\end{aligned}
$$

From the condition (3.3), the loss $H(B,\gamma)$ is strong convex. Hence we have

$$
\begin{aligned}
H(\hat{B},\hat{\gamma})\geq\ & H(B^*,\gamma^*)+\langle\nabla H_B(B^*,\gamma^*),\hat{B}-B^*\rangle+\langle\nabla H_\gamma(B^*,\gamma^*),\hat{\gamma}-\gamma^*\rangle\\
&+\sum_{i=1}^n K\langle X_i,\hat{B}-B^*\rangle^2+(\hat{\gamma}-\gamma^*)^T\sum_{i=1}^n Kz_iz_i^T(\hat{\gamma}-\gamma^*)\\
&+2\sum_{i=1}^n\langle X_i,\hat{B}-B^*\rangle\langle Kz_i,\hat{\gamma}-\gamma^*\rangle\\
\geq\ & H(B^*,\gamma^*)+\langle\nabla H_B(B^*,\gamma^*),\hat{B}-B^*\rangle+\langle\nabla H_\gamma(B^*,\gamma^*),\hat{\gamma}-\gamma^*\rangle\\
&+\frac{n}{2}\min\{\underline{C}_X,\underline{C}_Z\}(||\hat{B}-B^*||_F^2+||\hat{\gamma}-\gamma^*||^2). \qquad (A.24)
\end{aligned}
$$

The definition of $\hat{B},\hat{\gamma}$ gives us

$$
\begin{aligned}
&H(\hat{\beta},\hat{\gamma})+\lambda_1||\hat{B}||_*+\lambda_2||Cvec(\hat{B})||_1+\lambda_3||\hat{\gamma}||_1+\lambda_4||A_p\hat{\gamma}||_1\\
\leq& H(\beta^*,\gamma^*)+\lambda_1||B^*||_*+\lambda_2||Cvec(B^*)||_1+\lambda_3||\gamma^*||_1+\lambda_4||A_p\gamma^*||_1. \qquad (A.25)
\end{aligned}
$$

By Cauchy's inequality and Condition 8, we have

$$
\begin{aligned}
&\lambda_1||B^*||_*+\lambda_2||Cvec(B^*)||_1+\lambda_3||\gamma^*||_1+\lambda_4||A_p\gamma^*||_1\\
\leq&\lambda_1(r^*\|B^*\|_2)+\lambda_2(||Cvec(B^*)||_\infty s_1^*)+\lambda_3(s_2^*\|\gamma^*||_\infty)+\lambda_4(s_3^*\|A_p\gamma^*||_\infty)\\
\leq&4\sqrt{n}.
\end{aligned}
$$

Combing (A.23), (A.24) and (A.25) yields

$$
\begin{aligned}
&\frac{n}{2}\min\{\underline{C}_X,\underline{C}_Z\}(||\hat{B}-B^*||_F^2+||\hat{\gamma}-\gamma^*||^2)\\
\leq&|\langle\nabla H_B(B^*,\gamma^*),\hat{B}-B^*\rangle|+|\langle\nabla H_\gamma(B^*,\gamma^*),\hat{\gamma}-\gamma^*\rangle|+\lambda_1||B^*||_*\\
&+\lambda_2||Cvec(B^*)||_1+\lambda_3||\gamma^*||_1+\lambda_4||A_p\gamma^*||_1\\
\leq&|\langle\nabla H_B(B^*,\gamma^*),\hat{B}-B^*\rangle|+|\langle\nabla H_\gamma(B^*,\gamma^*),\hat{\gamma}-\gamma^*\rangle|+4\sqrt{n}. \qquad (A.26)
\end{aligned}
$$

Let $\varphi(B^*,\gamma^*)=\frac{e^{\langle X_i,B^*\rangle+\langle z_i,\gamma^*\rangle}}{1+e^{\langle X_i,B^*\rangle+\langle z_i,\gamma^*\rangle}}-y_i$. The first-order gradient of loss function is given as

$$
\nabla H_B(B^*,\gamma^*)=\sum_{i=1}^n X_i\varphi(B^*,\gamma^*),\ \nabla H_\gamma(B^*,\gamma^*)=\sum_{i=1}^n z_i\varphi(B^*,\gamma^*).
$$

Note that $|\varphi(B^*,\gamma^*)|\leq 1$, we obtain

$$
|\langle\nabla H_B(B^*,\gamma^*),\hat{B}-B^*\rangle|+|\langle\nabla H_\gamma(B^*,\gamma^*),\hat{\gamma}-\gamma^*\rangle|
$$

$$\begin{aligned}\le & |\langle \sum_{i=1}^{n} X_i, \hat{B} - B^*\rangle| + |\langle \sum_{i=1}^{n} z_i, \hat{\gamma} - \gamma^*\rangle| \\ \le & \sqrt{n}\bar{C}_X \|\hat{B} - B^*\|_F + \sqrt{n}\bar{C}_Z \|\hat{\gamma} - \gamma^*\|_2.\end{aligned}$$

(A.26) leads to

$$\begin{aligned}& \frac{n}{2}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2) \\ \le & \frac{\bar{C}_X}{\min\{\underline{C}_X, \underline{C}_Z\}} + \frac{\bar{C}_Z}{\min\{\underline{C}_X, \underline{C}_Z\}} + 4\sqrt{n}.\end{aligned}$$

We obtain

$$\begin{aligned}||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||_2^2 &\le \frac{1}{n}\Big(\frac{4\bar{C}_X}{\min^2\{\underline{C}_X, \underline{C}_Z\}} + \frac{4\bar{C}_Z}{\min^2\{\underline{C}_X, \underline{C}_Z\}}\Big) + 4/\sqrt{n} \\ &\le c^2.\end{aligned}$$

We have proved that $||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2$ has a upper bound. Next we will prove that the estimation error tends to zero. By (A.24) we know

$$\begin{aligned}& \frac{1}{2}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B} - B^*||_F^2 + ||\hat{\gamma} - \gamma^*||^2) \\ \le & \frac{1}{n}|\langle \nabla H_B(B^*, \gamma^*), \hat{B} - B^*\rangle + \langle \nabla H_\gamma(B^*, \gamma^*), \hat{\gamma} - \gamma^*\rangle| \\ & + \frac{\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1}{n}. \end{aligned} \tag{A.27}$$

For simplicity, denote $U = \hat{B} - B^*, v = \hat{\gamma} - \gamma^*$. We estimate the upper bound $\frac{1}{n}|\langle \nabla H_B(B^*, \gamma^*), \hat{B} - B^*\rangle + \langle \nabla H_\gamma(B^*, \gamma^*), \hat{\gamma} - \gamma^*\rangle|$.

$$\begin{aligned}& \sup_{\|U\|_F^2 + \|v\|^2 \le c} \frac{1}{n}|\langle \nabla H_B(B^*, \gamma^*), \hat{B} - B^*\rangle + \langle \nabla H_\gamma(B^*, \gamma^*), \hat{\gamma} - \gamma^*\rangle| \\ & = \sup_{\|U\|_F^2 + \|v\|^2 \le c} \frac{1}{n}|\langle \nabla H_B(B^*, \gamma^*), U\rangle + \langle \nabla H_\gamma(B^*, \gamma^*), v\rangle| \\ \le & \frac{2}{n} \sup_{\|U\|_F^2 + \|v\|^2 \le c} (\langle \sum_{i=1}^{n} X_i\varphi(B^*, \gamma^*), U\rangle)^2 \\ & + \frac{2}{n} \sup_{\|U\|_F^2 + \|v\|^2 \le c} (\langle \sum_{i=1}^{n} z_i\varphi(B^*, \gamma^*), v\rangle)^2 \\ \le & \sup_{\|U\|_F^2 + \|v\|^2 \le c} \frac{2}{n} \sum_{i=1}^{n} vec(U)^T vec(X_i) vec(X_i)^T vec(U) \\ & + \sup_{\|U\|_F^2 + \|v\|^2 \le c} \frac{2}{n} \sum_{i=1}^{n} v^T z_i z_i^T v.\end{aligned}$$

Condition 7 shows that the $vec(X_i)$ and z_i obey the sub-Guassian distributions. Hence the $vec(X_i)vec(X_i)^T$ and $z_iz_i^T$ satisfy the sub-exponential distributions. Applying Bernstein inequality for every $t \geq 0$ yields

$$\begin{aligned}&P\Big(\frac{1}{n}|\sum_{i=1}^{n} vec(U)^T vec(X_i)vec(X_i)^T vec(U)| > t\Big)\\&\leq 2exp\Big(-d\min\{\frac{nt^2}{\|X_i\|_{\psi_2}^4\|U\|_F^2}, \frac{nt}{\|X_i\|_{\psi_2}^2\|U\|_F}\}\Big).\end{aligned}$$

Taking $t = \sqrt{\frac{k^2\|U\|_F^2 mq}{n}}$ implies

$$\begin{aligned}&P\Big(\sup_{\|U\|_F^2+\|v\|^2\leq c}\frac{1}{n}|\sum_{i=1}^{n} vec(U)^T vec(X_i)vec(X_i)^T vec(U)| > \sqrt{\frac{k^2\|U\|_F^2 mq}{n}}\Big)\\&\leq c_1 exp(-c_2 mq),\end{aligned}$$

where $k = \max\{\|X_i\|_{\psi_2}^2, \|z_i\|_{\psi_2}^2\}$, c_1, c_2 are positive constants. Similarly,

$$P\Big(\sup_{\|U\|_F^2+\|v\|^2\leq c}\frac{1}{n}|\sum_{i=1}^{n} v^T z_i z_i^T v| > \sqrt{\frac{k^2\|v\|_2^2 p}{n}}\Big) \leq c_3 exp(-c_4 p),$$

where c_3, c_4 are positive constants. The (A.27) follows that

$$\begin{aligned}&\frac{1}{2}\min\{\underline{C}_X, \underline{C}_Z\}(||\hat{B}-B^*||_F^2 + ||\hat{\gamma}-\gamma^*||^2)\\
\leq&\frac{1}{n}|\langle\nabla H_B(B^*,\gamma^*), \hat{B}-B^*\rangle + \langle\nabla H_\gamma(B^*,\gamma^*), \hat{\gamma}-\gamma^*\rangle|\\
&+\frac{\lambda_1||B^*||_* + \lambda_2||Cvec(B^*)||_1 + \lambda_3||\gamma^*||_1 + \lambda_4||A_p\gamma^*||_1}{n}\\
\leq&\frac{1}{n}|\langle\nabla H_B(B^*,\gamma^*), \hat{B}-B^*\rangle + \langle\nabla H_\gamma(B^*,\gamma^*), \hat{\gamma}-\gamma^*\rangle|\\
&+\frac{\lambda_1 r^*\|B^*\|_2 + \lambda_2 s_1^*||Cvec(B^*)||_\infty + \lambda_3 s_2^*||\gamma^*||_\infty + \lambda_4 s_3^*\|A_p\gamma^*||_\infty}{n}\\
\leq&2\sqrt{\frac{k^2mq}{n}}\|\hat{B}-B^*\|_F + 2\sqrt{\frac{k^2p}{n}}\|\hat{\gamma}-\gamma^*\|_2\\
&+\frac{\lambda_1 r^*\|B^*\|_2 + \lambda_2 s_1^*||Cvec(B^*)||_\infty + \lambda_3 s_2^*||\gamma^*||_\infty + \lambda_4 s_3^*\|A_p\gamma^*||_\infty}{n}.\end{aligned}$$

Due to

$$\begin{aligned}2\sqrt{\frac{k^2mq}{n}}\|\hat{B}-B^*\|_F &= 2\cdot 2\sqrt{\frac{k^2mq}{n\min\{\underline{C}_X, \underline{C}_Z\}}}\frac{\sqrt{\min\{\underline{C}_X, \underline{C}_Z\}}}{2}\|\hat{B}-B^*\|_F\\
&\leq \frac{4k^2mq}{n\min\{\underline{C}_X, \underline{C}_Z\}} + \frac{1}{4}\min\{C_X, C_Z\}\|\hat{B}-B^*\|_F^2\end{aligned}$$

and

$$2\sqrt{\frac{k^2p}{n}}\|\hat{\gamma}-\gamma^*\|_2 = 2\cdot 2\sqrt{\frac{k^2p}{n\min\{\underline{C}_X,\underline{C}_Z\}}}\,\frac{\sqrt{\min\{\underline{C}_X,\underline{C}_Z\}}}{2}\|\hat{\gamma}-\gamma^*\|_2$$
$$\le \frac{4k^2p}{n\min\{\underline{C}_X,\underline{C}_Z\}} + \frac{1}{4}\min\{C_X,C_Z\}\|\hat{\gamma}-\gamma^*\|_2^2,$$

we have

$$||\hat{B}-B^*||_F^2 + ||\hat{\gamma}-\gamma^*||_2^2 \le \frac{16k^2(mp+q)}{n\min^2\{\underline{C}_X,\underline{C}_Z\}}$$
$$+\frac{4(\lambda_1 r^*\|B^*\|_2 + \lambda_2 s_1^*||Cvec(B^*)||_\infty + \lambda_3 s_2^*\|\gamma^*\|_\infty + \lambda_4 s_3^*\|A_p\gamma^*\|_\infty)}{n\min\{\underline{C}_X,\underline{C}_Z\}}$$

with probability at least $1-c_1e^{-c_2mq}-c_3e^{-c_4p}$. Further, under the Condition 8 and assumption for $\max\{mq,p\}/n \to 0$ as $n\to\infty$, it follows that $||\hat{B}-B^*||_F^2 + ||\hat{\gamma}-\gamma^*||_2^2$ converges in probability to zero. □

References

[1] Bahmani, S., Raj, B. and Boufounos, P. T. (2013). Greedy sparsity-constrained optimization. *Journal of Machine Learning Research* **14** 807–841. MR3049490

[2] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer. MR2807761

[3] Cai, J.-F., Candès, E. J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* **20** 1956–1982. MR2600248

[4] Chen, C., He, B., Ye, Y. and Yuan, X. (2016). The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming* **155** 57–79. MR3439797

[5] Chen, L., Sun, D. and Toh, K.-C. (2017). An efficient inexact symmetric Gauss–Seidel based majorized ADMM for high-dimensional convex composite conic programming. *Mathematical Programming* **161** 237-270. MR3592778

[6] Chung, C. F. and Agterberg, F. P. (1988). *Poisson regression analysis and its application*. Springer, Dordrecht.

[7] Dontchev, A. L. and Rockafellar, R. T. (2009). *Implicit functions and solution mappings*. Springer. MR2515104

[8] Elsener, A. and van de Geer, S. (2018). Robust low-rank matrix estimation. *Annals of Statistics* **46** 3481–3509. MR3852659

[9] Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *Annals of Statistics* **42** 324–351. MR3189488

[10] Fan, J., Gong, W. and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics* **212** 177–202. MR3994013

[11] Fan, J., Kong, L., Wang, L. and Xiu, N. (2018). Variable selection in sparse regression with quadratic measurements. *Statistica Sinica* **28** 1157–1178. MR3820999

[12] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[13] Fanaeet, H. and Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* **2** 113–127.

[14] Fazel, M., Pong, T. K., Sun, D. and Tseng, P. (2013). Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications* **34** 946–977. MR3073649

[15] Feng, Y. and Wang, S. (2017). A forecast for bicycle rental demand based on random forests and multiple linear regression. *Annual acis international conference on computer and information science* 101–105.

[16] Friedman, J. H., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1** 302–332. MR2415737

[17] Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications* **2** 17–40.

[18] Ghaoui, L. E., Viallon, V. and Rabbani, T. (2010). Safe feature elimination for the lasso and sparse supervised learning problems. *Pacific Journal of Optimization* **8** 667-698. MR3026449

[19] Grippo, L. and Sciandrone, M. (2000). On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations Research Letters* **26** 127–136. MR1746833

[20] Han, D., Sun, D. and Zhang, L. (2017). Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research* **43** 622–637. MR3801109

[21] Hestenes, M. R. (1969). Multiplier and gradient methods. *Journal of Optimization Theory and Applications* **4** 303–320. MR0271809

[22] Hung, H. and Wang, C. (2012). Matrix variate logistic regression model with application to EEG data. *Biostatistics* **14** 189–202.

[23] Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M. and Klein, M. (2002). *Logistic regression.*

[24] Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics* **5** 1161–1183. MR2842903

[25] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics* **39** 2302–2329. MR2906869

[26] Li, X., Mo, L., Yuan, X. and Zhang, J. (2014). Linearized alternating direction method of multipliers for sparse group and fused lasso models. *Computational Statistics and Data Analysis* **79** 203–221. MR3227997

[27] Li, X., Sun, D. and Toh, K.-C. (2018). On efficiently solving the sub-

problems of a level-set method for fused lasso problems. *SIAM Journal on Optimization* **28** 1842–1866. MR3816188

[28] Li, Y., Nan, B. and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240

[29] Liu, J., Ji, S. and Ye, J. (2013). SLEP: Sparse learning with efficient projections.

[30] Lu, Z., Monteiro, R. D. C. and Yuan, M. (2012). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming* **131** 163–194. MR2886145

[31] Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 53–71. MR2412631

[32] Moreau, J. J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes Rendus Hebdomadaires Des Sances De Lacadmie Des Sciences* **255** 2897–2899. MR0144188

[33] Moreau, J. J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France* **93** 273–299. MR0201952

[34] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics* **39** 1069–1097. MR2816348

[35] Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media. MR2244940

[36] Obozinski, B. Y. G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high dimensional multivariate regression. *Annals of Statistics* **39** 1–47. MR2797839

[37] Ortega, J. M. and Rheinboldt, W. C. (1967). Monotone iterations for nonlinear equations with application to Gauss-Seidel methods. *SIAM Journal on Numerical Analysis* **4** 171–190. MR0215487

[38] Plan, Y. and Vershynin, R. (2012). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory* **59** 482–494. MR3008160

[39] Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review* **52** 471–501. MR2680543

[40] Rockafellar, R. T. (1970). *Convex analysis*. Princeton University Press. MR0274683

[41] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*. Princeton University Press.

[42] Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Annals of Statistics* **39** 887–930. MR2816342

[43] Tian, G. L., Tang, M. L., Fang, H. B. and Tan, M. (2008). Efficient methods for estimating constrained parameters with applications to regularized (lasso) logistic regression. *Computational Statistics and Data Analysis* **52** 3528–3542. MR2427362

[44] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso.

Journal of the Royal Statistical Society: Series B (Methodological) **58** 267–288. MR1379242

[45] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108. MR2136641

[46] Usui, M., Niki, H. and Kohno, T. (1994). Adaptive Gauss-Seidel method for linear systems. *International Journal of Computer Mathematics* **51** 119–125. MR2577778

[47] Wahltinez, O., Lee, M., Erlinger, A., Daswani, M., Yawalkar, P., Murphy, K. and Brenner, M. (2020). COVID-19 Open-Data: curating a fine-grained, global-scale data repository for SARS-CoV-2.

[48] Wang, L., You, Y. and Lian, H. (2013). A simple and efficient algorithm for fused lasso signal approximator with convex loss function. *Computational Statistics* **28** 1699–1714. MR3120835

[49] Zhang, Y., Zhang, N., Sun, D. and Toh, K.-C. (2018). An efficient Hessian based algorithm for solving large-scale sparse group Lasso problems. *Mathematical Programming* **179** 1–41. MR4050140

[50] Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 463–483. MR3164874

[51] Zhou, Z. and So, A. M.-C. (2017). A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming* **165** 689–728. MR3707378

[52] Zhu, Y. (2017). An augmented ADMM algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics* **26** 195–204. MR3610420

[53] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320. MR2137327

[54] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* **37** 1733–1751. MR2533470