Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 17, No. 1, pp. 1–22, 2021

SINGULAR VALUE SCREENING RULES FOR THE NUCLEAR NORM REGULARIZED MULTIVARIATE LINEAR REGRESSION*

Pan Shang† and Lingchen Kong

Abstract: Multivariate linear regression is a major concern in statistics and machine learning. To capture the low rank structure in high-dimensional data sets, one popular model is the nuclear norm regularized multivariate linear regression (NMLR). Recently, there are some screening rules for LASSO that can eliminate inactive features in data sets under different tuning parameters. To the best of our knowledge, there are no screening rules for NMLR in the view of optimization. To establish such rules, we study the duality theory of NMLR and find out that its dual solution is a projection on a compact and convex set. According to properties of the projection operator, we establish our screening rules, which identify inactive singular values and estimate the maximal rank of the solution. In addition, we propose an alternating direction method of multipliers (ADMM) to solve NMLR. The numerical experiments on simulation and real data sets illustrate the efficiency of our algorithm.

Key words: *screening rule, singular value, nuclear norm regularized multivariate linear regression, duality theory, projection operator*

Mathematics Subject Classification: *90C46, 90C06, 90C25, 62J99*

1 Introduction

High-dimensional multivariate linear regression is widely used in many research areas, such as environmental science (Abyaneh [1]), gene expression (Peng et al. [23]), operations research (Noorossana et al. [21]) and so on. For example, Peng et al. [23] measured the influence of DNA copy number alterations on RNA transcript levels in a breast cancer study. This study includes 172 samples, 384 DNA copy number and 654 breast cancer related RNA expressions for every sample. Here, the predictors are 384 DNA copy number and the responses are 654 breast cancer related RNA expressions. Thus, the prediction matrix is 172 by 384 and the response matrix is 172 by 654. To explore this influence, a direct way is the multivariate linear regression. Note that the sample size is less than the number of predictors or responses. So, the data is high-dimensional. For this type data, one common assumption of the coefficient matrix is low rank. However, optimization problems with the

*This work was supported by the National Natural Science Foundation of China (12071022).
†Corresponding author.

low rank constraint are NP-hard. The regularization technique is always used to deal with these problems, and one popular regularizer is the nuclear norm (see, e.g., Agarwal et al. [2], Negahban et al. [20] and Yuan et al. [31]). Therefore, our concern is the nuclear norm regularized multivariate linear regression (NMLR).

Recently, there are some screening rules for LASSO with the help of optimization techniques, which identify inactive predictors under different tuning parameters. See, e.g., Fan et al. [9], Ghaoui et al. [8], Tibshirani et al. [26], Wang et al. [28], Ndiaye et al. [19], Kuang et al. [15], Xiang et al. [30], Lee et al. [16], Pan and Xu [22]. For example, Fan et al. [9] proposed the sure independence screening (SIS), which reduces dimensionality of predictors below sample size. Ghaoui et al. [8] constructed SAFE rules to eliminate inactive predictors, which are based on the duality theorem in optimization. The SAFE rules never remove active predictors. Tibshirani et al. [26] proposed strong rules for discarding inactive predictors under the assumption of the unit slope bound. The strong rules screen out far more predictors than SAFE rules in practice and can be more efficient by checking Karush-Kuhn-Tucker (KKT) conditions for any predictor. Wang et al. [28] built the dual polytope projection (DPP) and the enhanced version EDPP to discard inactive predictors. They showed that EDPP had a better performance in screening out inactive predictors than SAFE rules and strong rules. Ndiaye et al. [19] built up statics and dynamic gap safe screening rules, which are based on the gap between feasible points of LASSO and its dual problem. As we all know, the sparsity means many elements of the coefficient vector are zero, which implies lots of the predictors are inactive. Therefore, under different tuning parameters, these screening rules provide the estimation of the sparsity of the coefficient vector. By analyzing the above results, we know that optimization techniques play important roles in identifying inactive predictors for LASSO. This opens a hope that we may build up screening rules for NMLR from the point of view of optimization. However, the low rank matrix does not mean the sparsity of its elements, but the sparsity of its singular value vector. One nature question is: *how to establish the screening rules for NMLR?*

This paper will deal with this problem and give an affirmative answer. In order to do so, we present the dual problem of NMLR and find that the dual solution is the projection on a compact and convex set. With the help of the strong duality theorem, we present that inactive singular values of the solution of NMLR can be identified by its dual solution. However, the dual solution may be complex to be computed for every tuning parameter. Based on the nonexpansivity of the projection operator, we give an estimate set for the dual solution of NMLR. Therefore, we get a singular value screening rule SSR, which identifies inactive singular values and estimate the maximal rank of the solution of NMLR, under different tuning parameters. According to the firm nonexpansivity of the projection operator, we continue to get the enhance version SSR+, which behaves better than SSR surely. In addition, we propose an alternating direction method of multipliers (ADMM) to solve NMLR. To illustrate the efficiency of our algorithm, we implement some numerical experiments on simulation and real data sets. By comparing ADMM with existing solvers SLEP and CVX, we verify that ADMM is an efficient way to solve NMLR.

The rest of this paper is organized as follows. In Section 2, we review some basic concepts and results. Section 3 presents the model analysis of NMLR and its duality theory. In Section 4, we propose two singular value screening rules based on properties of the projection operator. In Section 5, we propose an alternating direction method of multipliers(ADMM)

to solve NMLR. To verify the efficiency of ADMM, we also implement some numerical experiments. Some conclusion remarks are given in Section 6.

Notations: Let $M \in \mathbb{R}^{p\times q}$ be any matrix. Suppose M has a singular value decomposition with nondecreasing singular values $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_r(M) \geq 0$, where r=min$\{p, q\}$ and it is used throughout this paper. There are some norms related to M and these definitions are used throughout the paper. The Frobenius norm $\|\cdot\|_F$ is defined as $\|M\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} M_{ij}^2} = \sqrt{\mathrm{tr}(\mathrm{M^T M})} = \sqrt{\sigma_1(M)^2 + \cdots + \sigma_r(M)^2}$. The nuclear norm $\|\cdot\|_*$ is the sum of singular values, i.e., $\|M\|_* = \sum_{i=1}^{r}\sigma_i(M)$. The spectral norm $\|\cdot\|_2$ is the largest singular value, i.e., $\|M\|_2 = \sigma_1(M)$. For any vector $\mathbf{x} \in \mathbb{R}^n$, the 2-norm $\|\cdot\|_2$ is defined as $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

2 Preliminaries

In this section, we give some basic concepts and results in optimization. The following definitions and results are from Rockafellar [24].

Definition 2.1. Let $f : \mathbb{R}^{p\times q} \to \mathbb{R}$. For any $M \in \mathbb{R}^{p\times q}$, the conjugate function $f^* : \mathbb{R}^{p\times q} \to \mathbb{R}$ is defined as

$$f^*(M) = \sup_{N\in dom(f)} \{\langle M, N\rangle - f(N)\}.$$

If $f(M) = \|M\|_*$, we can get

$$f^*(M) = \sup_{N\in\mathbb{R}^{p\times q}} \{\langle M, N\rangle - \|N\|_*\} = \delta_{\|\cdot\|_2\leq 1}(M),$$

where $\delta_{\|\cdot\|_2\leq 1}(M)$ is an indicator function defined as

$$\delta_{\|\cdot\|_2\leq 1}(M) = \begin{cases} 0, & \|M\|_2 \leq 1 \\ +\infty, & \|M\|_2 > 1. \end{cases}$$

If $f(M) = \frac{1}{2}\|M\|_F^2$,

$$f^*(M) = \sup_{N\in\mathbb{R}^{p\times q}} \left\{\langle M, N\rangle - \frac{1}{2}\|N\|_F^2\right\} = \frac{1}{2}\|M\|_F^2.$$

Definition 2.2. For an arbitrary vector ω and a convex set Ω, the projection operator $P_\Omega(\cdot)$ is defined as

$$P_\Omega(\omega) = \underset{\mu\in\Omega}{\mathrm{argmin}} \ \frac{1}{2}\|\mu - \omega\|_2^2.$$

Here are some basic properties of projection operator.

Lemma 2.3. *Let Ω be any compact and convex set, then the projection operator on Ω is*

(1) *nonexpansive, i.e.,*

$$\|P_\Omega(\omega_2) - P_\Omega(\omega_1)\| \leq \|\omega_2 - \omega_1\|, \quad \forall \omega_1, \omega_2 \in \Omega.$$

(2) *firmly nonexpansive, i.e.,*

$$\|P_{\Omega}(\omega_1) - P_{\Omega}(\omega_2)\|^2 + \|(I - P_{\Omega})(\omega_1) - (I - P_{\Omega})(\omega_2)\|^2 \leq \|\omega_1 - \omega_2\|^2, \forall \omega_1, \omega_2 \in \Omega,$$

where I is the identity operator.

For singular values of matrixes, there are some basic inequalities (Roger [25]) that are useful for our paper.

Lemma 2.4. *Suppose $P, Q \in \mathbb{R}^{p\times q}$ and r=min$\{p, q\}$.* Two basic inequalities for singular values are showed as follows.

$$\sigma_{i+j-1}(P+Q) \leq \sigma_i(P) + \sigma_j(Q), \qquad i,j \in \{1,2,\ldots,r\}, i+j \leq r+1;$$

$$\sigma_{i+j-1}\left(PQ^T\right) \leq \sigma_i(P)\,\sigma_j(Q), \qquad i,j \in \{1,2,\ldots,r\}, i+j \leq r+1.$$

In particular,

$$\sigma_i(P+Q) \leq \sigma_i(P) + \sigma_1(Q), \qquad i \in \{1,2,\ldots,r\};$$

$$\sigma_1\left(PQ^T\right) \leq \sigma_1(P)\sigma_1(Q) = \|P\|_2\|Q\|_2.$$

3 Duality theory

In this section, we introduce the nuclear norm regularized multivariate linear regression (NMLR) and show its duality theory from the optimization perspective.

The nuclear norm regularized multivariate linear regression (NMLR, Yuan et al. [31]) is given as

$$\min_{B\in\mathbb{R}^{p\times q}} \left\{ \frac{1}{2}\|Y - XB\|_F^2 + \lambda\|B\|_* \right\}, \tag{3.1}$$

where $\lambda > 0$ is the tuning parameter. In this model, $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ is the prediction matrix and $Y = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^T \in \mathbb{R}^{n\times q}$ is the response matrix. The solution of NMLR relies on the choice of the tuning parameter λ, so we denote it as $B^*(\lambda)$.

Now, we consider the duality theory of NMLR. First, we rewrite NMLR as a constraint problem, which is

$$\begin{aligned} &\min_{B,A} \left\{ \lambda\|B\|_* + \frac{1}{2}\|A\|_F^2 \right\} \\ &s.t. \quad Y - XB - A = 0. \end{aligned} \tag{3.2}$$

Thus, we have the Lagrangian function of (3.2) as

$$L\left(B, A; \widetilde{C}\right) = \lambda\|B\|_* + \frac{1}{2}\|A\|_F^2 + \left\langle \widetilde{C}, Y - XB - A \right\rangle.$$

where $\widetilde{C} \in R^{n\times q}$ is a Lagrangian multiplier. We have the Lagrangian dual problem of (3.2) as

$$\max_{\widetilde{C}} \min_{B,A} \left\{ L\left(B, A; \widetilde{C}\right) \right\}.$$

It is not hard to yield the closed form of $\min_{B,A} L\left(B, A; \widetilde{C}\right)$ as follows.

$$\begin{aligned}\min_{B,A} L\left(B, A; \widetilde{C}\right) &= \min_{B,A}\left\{\lambda\|B\|_* + \frac{1}{2}\|A\|_F^2 + \left\langle \widetilde{C}, Y - XB - A\right\rangle\right\} \\ &= \min_{B}\left\{\lambda\|B\|_* - \left\langle X^T\widetilde{C}, B\right\rangle\right\} + \min_{A}\left\{\frac{1}{2}\|A\|_F^2 - \left\langle \widetilde{C}, A\right\rangle\right\} \\ &+ \left\langle \widetilde{C}, Y\right\rangle \\ &= -\delta_{\|\cdot\|_2\leq\lambda}\left(X^T\widetilde{C}\right) - \frac{1}{2}\|\widetilde{C}\|_F^2 + \left\langle \widetilde{C}, Y\right\rangle.\end{aligned}$$

The last equality is a direct result of the conjugate function. Thus, the dual problem of (3.2) is

$$\begin{aligned}&\max_{\widetilde{C}}\left\{-\frac{1}{2}\|\widetilde{C} - Y\|_F^2 + \frac{1}{2}\|Y\|_F^2\right\} \\ &s.t. \quad \|X^T\widetilde{C}\|_2 \leq \lambda.\end{aligned}$$

Taking $C = \frac{\widetilde{C}}{\lambda}$, the last equation is equivalent to

$$\begin{aligned}&-\min_{C}\left\{\frac{\lambda^2}{2}\left\|C - \frac{Y}{\lambda}\right\|_F^2 - \frac{1}{2}\|Y\|_F^2\right\} \\ &s.t. \quad \|X^TC\|_2 \leq 1.\end{aligned} \tag{3.3}$$

Denote the feasible area of (3.3) as

$$\Omega_D = \left\{C \middle| \|X^TC\|_2 \leq 1\right\}.$$

It is clear that Ω_D is a compact and convex set, and the solution of (3.3) is

$$C^*(\lambda) = P_{\Omega_D}\left(\frac{Y}{\lambda}\right), \tag{3.4}$$

where $P_{\Omega_D}(\cdot)$ denotes the projection operator on Ω_D. Note that Ω_D is not a polytope, which is different from LASSO in vector case, see, Wang et al. [28]. From the optimality conditions analysis, we have the Karush-Kuhn-Tucker (KKT) system of (3.2) and (3.3)

$$\begin{cases}X^TC \in \partial\|B\|_*, \\ A = \lambda C, \\ Y - XB - A = 0.\end{cases} \tag{3.5}$$

If a pair $(B^*(\lambda), A^*(\lambda), C^*(\lambda))$ satisfies the KKT system, it is called the KKT point of (3.2) and (3.3). Based on the convex optimization analysis, it holds the strong duality theorem.

Theorem 3.1 (Strong duality theorem)**.** *The linear constraint optimization problem (3.2) satisfies Slater's constraint qualification and there is a KKT point $(C^*(\lambda), B^*(\lambda), A^*(\lambda))$ such that the optimal values of (3.2) and (3.3) are equal, i.e.,*

$$\lambda\|B^*(\lambda)\|_* + \frac{1}{2}\|A^*(\lambda)\|_F^2 = -\left(\frac{\lambda^2}{2}\left\|C^*(\lambda) - \frac{Y}{\lambda}\right\|_F^2 - \frac{1}{2}\|Y\|_F^2\right).$$

Here, $(B^*(\lambda), A^*(\lambda))$ is the solution of (3.2) and $C^*(\lambda)$ is the solution of (3.3).

Proof. Now we discuss the relationship between (3.2) and (3.3). The objective function of (3.2) is

$$f := \left\{ \lambda \|B\|_* + \frac{1}{2}\|A\|_F^2 \right\}$$

and the feasible area $S := \left\{ (B, A) \middle| Y - XB - A = 0 \right\}$. For convex optimization problems with linear constraints, there is an important constraint qualification named Slater's constraint qualification (Slater's CQ). If a convex optimization problem satisfies Slater's CQ, it follows from Rockafellar [24] that the solutions of primal and dual problems are KKT points.

Slater's CQ: There exists $\theta \in \text{ri}(\text{dom } (f)) \bigcap S$, where f is the objective function and S is the feasible area of optimization problem.

It is clear that there exist

$$B = 0, A = Y \text{ such that } Y - XB - A = 0,$$

which means that (3.2) satisfies Slater's CQ. Because Ω_D is a compact and convex set and $C^*(\lambda) = P_{\Omega_D}(\frac{Y}{\lambda})$, it is sure that (3.3) have a solution. By solving (3.4) under $C = C^*(\lambda)$, we obtain the solution of (3.2). So, based on the Rockafellar [24], the strong duality theorem holds on (3.2) and (3.3). □

4 Singular value screening rules

In this section, we give two singular value screening rules for NMLR, which can identify inactive singular values and estimate the maximal rank of the solution of NMLR.

For NMLR, we already know that $B^*(\lambda) = 0$ if λ is sufficiently large. The next proposition gives the lower bound of the tuning parameter λ which guarantees $B^*(\lambda) = 0$.

Proposition 4.1. *$B^*(\lambda) = 0$ is the solution of NMLR if and only if $\lambda \geq \lambda_{max} := \|X^T Y\|_2$.*

Proof. We first prove the "only if" part. Based on (3.5), it is obvious that the solution of (3.3) is $C^*(\lambda) = \frac{Y}{\lambda}$, if $B^*(\lambda) = 0$. It means $\frac{Y}{\lambda} \in \Omega_D$ and $\|X^T \frac{Y}{\lambda}\|_2 \leq 1$. That is $\|X^T Y\|_2 \leq \lambda$. Therefore, $\lambda \geq \lambda_{max} = \|X^T Y\|_2$.

Now we prove the "if" part. If $\lambda \geq \lambda_{max}$, we can get $C = \frac{Y}{\lambda} \in \Omega_D$. Under the fact that $C^*(\lambda) = P_{\Omega_D}\left(\frac{Y}{\lambda}\right)$, the solution of (3) is $C^*(\lambda) = \frac{Y}{\lambda}$. According to the $A^*(\lambda) = \lambda C^*(\lambda)$ in (3.5), we have $A^*(\lambda) = Y$. By Theorem 3.1, we get

$$\frac{1}{2}\|Y\|_F^2 + \lambda \|B^*(\lambda)\|_* = \frac{1}{2}\|Y\|_F^2.$$

This yields $\|B^*(\lambda)\|_* = 0$, which implies $B^*(\lambda) = 0$. □

From Proposition 4.1, we focus on the case that $\lambda \in (0, \lambda_{max})$ in the rest of this paper. Before presenting the detailed process of our screening rules, we get a relationship of the rank of $B^*(\lambda)$ and $X^T Y$ in some special cases, where we do not need to apply any screening rule. That is, when all singular values of $X^T Y$ are equal to a certain number, the rank of $B^*(\lambda)$ can be directly determined.

Proposition 4.2. 1. *If $X^TY = 0$, then rank$(B^*(\lambda)) = 0$ holds for any $\lambda > 0$.*

2. *If there exists $\alpha > 0$ such that $\sigma_i(X^TY) = \alpha$ holds for any $i \in \{1, 2, \ldots, r\}$, then rank$(B^*(\lambda)) = r$ when $\lambda \in (0, \lambda_{max})$.*

Proof. Case 1: $X^TY = 0$. In this case, $\lambda_{max} = 0$ and $B^*(\lambda) = 0$ for any $\lambda > 0$.
Case 2: All singular values of X^TY are equal to a nonzero number α, which means rank$(X^TY) = r$. In this case, $\lambda_{max} = \alpha$ and $C^*(\lambda_{max}) = \frac{Y}{\alpha} \in \Omega_D$. For any $0 < \lambda < \alpha$,

$$C^*(\lambda) = P_{\Omega_D}\left(\frac{Y}{\lambda}\right) = \frac{Y}{\alpha}.$$

Replacing it into (3.5), we know that

$$X^TY = X^T\left(XB^*(\lambda) + \lambda\frac{Y}{\alpha}\right),$$

which leads to

$$\left(1 - \frac{\lambda}{\alpha}\right) X^TY = X^T\left(XB^*(\lambda)\right).$$

Hence, rank$(X^TXB^*(\lambda))$=rank$(X^TY) = r$. By using the fact that

$$\text{rank}(X^TXB^*(\lambda)) \leq \text{rank}(B^*(\lambda)) \leq r,$$

we know that rank$(B^*(\lambda)) = r$. Therefore, the result is proved. □

Proposition 4.2 illustrates that we do not need to identify inactive singular values of $B^*(\lambda)$ in some cases, where all singular values of X^TY are equal to a certain number. In the following parts, we discuss the case that X^TY has at least two different singular values.

According to the $X^TC \in \partial\|B\|_*$ in (3.5) and Theorem 3.1, we easily obtain a sufficient condition for identifying inactive singular values of $B^*(\lambda)$.

Theorem 4.3. *Let $i \in \{1, 2, \ldots, r\}$ and $\lambda \in (0, \lambda_{max})$. If $\sigma_i(X^TC^*(\lambda)) < 1$, then*

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i, \ldots, r\},$$

where $B^*(\lambda), C^*(\lambda)$ are solutions of (3.1) and (3.3), respectively.

Proof. Suppose the singular value decomposition of $B^*(\lambda)$ is

$$B^*(\lambda) = U\Sigma V^T,$$

where

$$\Sigma_{ij} = \begin{cases} 0, & i \neq j \\ \sigma_i(B^*(\lambda)), & i = j \end{cases}, i \in \{1, 2, \ldots, p\}, j \in \{1, 2, \ldots, q\}$$

with $\sigma_1(B^*(\lambda)) \geq \sigma_2(B^*(\lambda)) \geq \cdots \geq \sigma_r(B^*(\lambda)) \geq 0$ being the singular values of $B^*(\lambda)$ and $r = \min\{p, q\}$. Then, the subdifferential of $\|B^*(\lambda)\|_*$ (Watson [29]) is

$$\partial\|B^*(\lambda)\|_* = \left\{ UWV^T \middle| W \in R^{p\times q}, W_{ij} \in \begin{cases} \{0\}, & i \neq j \\ \{1\}, & i = j, \sigma_i(B^*(\lambda)) > 0 \\ [0, 1], & i = j, \sigma_i(B^*(\lambda)) = 0 \end{cases} \right\}.$$

For (3.5), we know that

$$X^T C^*(\lambda) \in \partial \|B^*(\lambda)\|_*.$$

Therefore, if

$$\sigma_i(X^T C^*(\lambda)) < 1,$$

then $\sigma_i(B^*(\lambda)) = 0$, which means $\sigma_j(B^*(\lambda)) = 0$ holds for any $j \in \{i, \ldots, r\}$. □

For any fixed tuning parameter, Theorem 4.3 implies that sample data and the solution of (3.3) can identify inactive singular values of the solution of NMLR. It is worth noting that the number of inactive singular values decides the rank of a matrix. Based on this, the rank of the solution of NMLR can be yielded.

However, the solution of (3.3) may be complex to be computed. It is fortunate that this solution is a projection on a compact and convex set. Hence, properties of the projection operator can help us to estimate the dual solution and establish screening rules for NMLR.

4.1 SSR

In this section, we estimate the dual solution of NMLR with the help of the nonexpansivity of the projection operator in Lemma 2.3. Then, we propose a singular value screening rule based on the estimate area of the dual solution and Theorem 4.3.

Theorem 4.4 (SSR)**.** *Let $\lambda_0 \in (0, \lambda_{max})$. Assume the solution $C^*(\lambda_0)$ of (3.3) is known. For any $i \in \{1, \ldots, r\}$, if $\lambda < \lambda_0$ and*

$$\lambda > \frac{\lambda_0 \|X\|_2 \|Y\|_F}{\lambda_0 - \lambda_0 \sigma_i\left(X^T C^*(\lambda_0)\right) + \|X\|_2 \|Y\|_F}, \tag{4.1}$$

we have

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i, \ldots, r\},$$

which leads to the solution of NMLR satisfies rank$(B^*(\lambda)) \leq i - 1$.

Proof. According to (3.4) and the nonexpansivity of $P_{\Omega_D}(\cdot)$ in Lemma 2.3, we know that

$$\|C^*(\lambda) - C^*(\lambda_0)\|_F \leq \left(\tfrac{1}{\lambda} - \tfrac{1}{\lambda_0}\right) \|Y\|_F.$$

Setting $\rho_1 := \left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) \|Y\|_F$, we define the set $\Omega_1 := \left\{C \middle| \ \|C - C^*(\lambda_0)\|_F \leq \rho_1\right\}$.

In order to prove the desired result, it is enough to consider $\sup\limits_{C \in \Omega_1} \left\{\sigma_i\left(X^T C\right)\right\}$ by Theorem 4.3. In fact, if $\sup\limits_{C \in \Omega_1} \left\{\sigma_i\left(X^T C\right)\right\} < 1$, $\sigma_i\left(C^*(\lambda)\right) < 1$ must hold, which leads to $\sigma_j\left(B^*(\lambda)\right) = 0, \forall j \in \{i, \ldots, r\}$.

Based on the definition of Ω_1, we know that

$$\begin{aligned} \sup_{C \in \Omega_1} \left\{\sigma_i\left(X^T C\right)\right\} &= \sup_{\|D\|_F \leq \rho_1} \left\{\sigma_i\left(X^T\left(C^*(\lambda_0) + D\right)\right)\right\} \\ &= \sup_{\|D\|_F \leq \rho_1} \left\{\sigma_i\left(X^T C^*(\lambda_0) + X^T D\right)\right\}. \end{aligned}$$

According to Lemma 2.4, we can get

$$\sigma_i\left(X^TC^*\left(\lambda_0\right)+X^TD\right)\leq\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\sigma_1\left(X^TD\right)$$

and

$$\sigma_1\left(X^TD\right)\leq\sigma_1(X)+\sigma_1(D)=\|X\|_2\|D\|_2.$$

Hence,

$$\begin{aligned}\sup_{C\in\Omega_1}\ \left\{\sigma_i\left(X^TC\right)\right\}&=\sup_{\|D\|_F\leq\rho_1}\ \left\{\sigma_i\left(X^TC^*\left(\lambda_0\right)+X^TD\right)\right\}\\&\leq\sup_{\|D\|_F\leq\rho_1}\ \left\{\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\sigma_1\left(X^TD\right)\right\}\\&\leq\sup_{\|D\|_F\leq\rho_1}\ \left\{\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\|X\|_2\|D\|_2\right\}\\&\leq\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\rho_1\|X\|_2.\end{aligned}$$

The last equality is obtained by the fact that $\|D\|_2\leq\|D\|_F\leq\rho_1$. Suppose $\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\rho_1\|X\|_2<1$, that is

$$\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)\leq1-\left(\frac{1}{\lambda}-\frac{1}{\lambda_0}\right)\|X\|_2\|Y\|_F. \tag{4.2}$$

We have $\sup_{C\in\Omega}\left\{\sigma_i\left(X^TC\right)\right\}<1$ and $\sigma_i\left(X^TC^*\left(\lambda\right)\right)<1$. Therefore, $\sigma_i\left(B^*(\lambda)\right)=0$, which implies that $\sigma_j\left(B^*(\lambda)\right)=0$ for any $j\in\{i,\ldots,r\}$.

To obtain the closed-form of λ , we multiply λ by both sides of (4.2) and get that

$$\lambda\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)\leq\lambda-\|X\|_2\|Y\|_F+\tfrac{\lambda}{\lambda_0}\|X\|_2\|Y\|_F.$$

By transforming all terms about λ to the right side of the last inequality, we get that

$$\begin{aligned}\|X\|_2\|Y\|_F&\leq\lambda-\lambda\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\frac{\lambda}{\lambda_0}\|X\|_2\|Y\|_F\\&=\left(1-\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\frac{1}{\lambda_0}\|X\|_2\|Y\|_F\right)\lambda,\end{aligned}$$

which is equivalent to

$$\lambda>\frac{\lambda_0\|X\|_2\|Y\|_F}{\lambda_0-\lambda_0\sigma_i\left(X^TC^*\left(\lambda_0\right)\right)+\|X\|_2\|Y\|_F}.$$

Therefore, the desired result follows. □

If λ_0 is set and the solution $C^*(\lambda_0)$ of (3.3) is easy to be solved, Theorem 4.4 claims that we can identify inactive singular values and estimate the maximal rank of the solution of NMLR under different tuning parameters. In general, $C^*(\lambda_0)$ may not be computed easily for a given λ_0. Fortunately, for $\lambda_0=\lambda_{max}=\|X^TY\|_2$, the solution $C^*(\lambda_0)$ equals to $\frac{Y}{\lambda_0}$ from the proof of Proposition 4.1.

Corollary 4.5. *Suppose X^TY has at least two different singular values. We define the sequence of tuning parameters $\{\lambda_i^{(1)}\}_{i=1}^r$ as*

$$\lambda_i^{(1)} = \frac{\|X^TY\|_2\|X\|_2\|Y\|_F}{\|X^TY\|_2 - \sigma_i\left(X^TY\right) + \|X\|_2\|Y\|_F}.$$

Let $i \in \{2,\ldots,r\}$. If $\lambda \in (\lambda_i^{(1)}, \lambda_{i-1}^{(1)}]$, then

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i,\ldots,r\},$$

which leads to the solution of NMLR satisfies rank$(B^*(\lambda)) \leq i-1$.

Proof. According to the definition in this corollary, we know that $\lambda_1 = \lambda_{max}$. The choice of λ in Theorem 4.4 should be satisfied that $\lambda < \lambda_{max}$, so we just discuss about the case that $i \geq 2$.

Let $\lambda_0 = \lambda_{max} = \|X^TY\|_2$. Replacing $C^*(\lambda_0) = \frac{Y}{\lambda_0}$ into (4.1), we know that

$$\begin{aligned}\lambda > & \frac{\lambda_0\|X\|_2\|Y\|_F}{\lambda_0 - \lambda_0\sigma_i\left(X^T\frac{Y}{\lambda_0}\right) + \|X\|_2\|Y\|_F} \\ & = \frac{\|X^TY\|_2\|X\|_2\|Y\|_F}{\|X^TY\|_2 - \sigma_i\left(X^TY\right) + \|X\|_2\|Y\|_F} \\ & = \lambda_i^{(1)}.\end{aligned}$$

Because $\sigma_i\left(X^TY\right) \leq \sigma_{i-1}\left(X^TY\right) \leq \|X^TY\|_2$, $\lambda_{i-1}^{(1)} \geq \lambda_i^{(1)}$ holds for any $i \in \{2,\ldots,r\}$.

From Theorem 4.4, if $\lambda > \lambda_i^{(1)}$,

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i,\ldots,r\}.$$

Similarly, if $\lambda > \lambda_{i-1}^{(1)}$,

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i-1,\ldots,r\}.$$

Combining these two results, we have

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i,\ldots,r\},$$

when $\lambda \in (\lambda_i, \lambda_{i-1}]$. Thus, the conclusion is proved. □

Corollary 4.5 is a special case of Theorem 4.4 when $\lambda_0 = \lambda_{max}$. This result shows that the inactive singular values of the solution of NMLR can be identified by the sample data.

From the proof of Theorem 4.4, we know

$$C^*(\lambda) \in \Omega_1 = \left\{C \middle| \|C - C^*(\lambda_0)\|_F \leq \rho_1\right\}.$$

The more accurate of Ω_1, the more accurate of the singular value screening rule. Therefore, the aim of the next part is to reach a more accurate estimate of $C^*(\lambda)$, which directly results in improvement consequences of SSR. A nature idea is to improve the result by using the other properties of the projection operator.

4.2 SSR+

In this section, we get another screening rule SSR+ based on the firm nonexpansivity of the projection operator. In order to obtain the results, we first give a lemma.

Lemma 4.6. *Let $\lambda_0 \in (0, \lambda_{max})$. Suppose the solution $C^*(\lambda_0)$ of (3.3) is known. Let $0 < \lambda < \lambda_0$. The dual solution $C^*(\lambda)$ can be estimated as*

$$C^*(\lambda) \in \Omega_2 \subseteq \Omega,$$

where $\Omega_2 := \left\{ C \middle| \left\| C - C^*(\lambda_0) - \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right\|_F \leq \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) \|Y\|_F \right\}$.

Proof. According to the firm nonexpansivity in Lemma 2.3, we have

$$\|C^*(\lambda) - C^*(\lambda_0)\|_F^2 + \left\| \frac{Y}{\lambda} - C^*(\lambda) - \frac{Y}{\lambda_0} + C^*(\lambda_0) \right\|_F^2 \leq \left\| \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\|_F^2. \tag{4.3}$$

Extending $\left\| \frac{Y}{\lambda} - C^*(\lambda) - \frac{Y}{\lambda_0} + C^*(\lambda_0) \right\|_F^2$, (4.3) is equivalent to

$$\begin{aligned} &\|C^*(\lambda) - C^*(\lambda_0)\|_F^2 + \left\| \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\|_F^2 \\ &- 2\left\langle C^*(\lambda) - C^*(\lambda_0), \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\rangle + \|C^*(\lambda) - C^*(\lambda_0)\|_F^2 \\ &\leq \left\| \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\|_F^2. \end{aligned}$$

This can be reformulated as $\|C^*(\lambda) - C^*(\lambda_0)\|_F^2 \leq \left\langle C^*(\lambda) - C^*(\lambda_0), \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\rangle$, which is equivalent to

$$\left\| C^*(\lambda) - C^*(\lambda_0) - \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right\|_F^2 \leq \frac{1}{4}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)^2 \|Y\|_F^2.$$

From the definition of Ω_2, we know that $C^*(\lambda) \in \Omega_2$.

For any $C \in \Omega_2$, we know that

$$\begin{aligned} \|C - C^*(\lambda_0)\|_F^2 &\leq \left\langle C - C^*(\lambda_0), \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\rangle \\ &\leq \|C - C^*(\lambda_0)\|_F \cdot \left\| \frac{Y}{\lambda} - \frac{Y}{\lambda_0} \right\| \\ &= \|C - C^*(\lambda_0)\|_F \left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) \|Y\|_F. \end{aligned}$$

The second inequality holds because of the Cauchy inequality. So, we have

$$\|C - C^*(\lambda_0)\|_F \leq \left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) \|Y\|_F,$$

which means $C \in \Omega_1$ and $\Omega_2 \subseteq \Omega$. □

By using the similar idea in Theorem 4.4, we get the following result.

Theorem 4.7 (SSR+)**.** *Let $\lambda_0 \in (0, \lambda_{max})$. Assume the solution $C^*(\lambda_0)$ of (3) is known. For any $i \in \{1, 2, \ldots, r\}$, if $\lambda < \lambda_0$ and*

$$\lambda > \frac{\lambda_0 \|X\|_2 \|Y\|_F}{2\lambda_0 - 2\lambda_0 \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \|X\|_2 \|Y\|_F},$$

we have

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i, \ldots, r\},$$

which leads to the solution of NMLR satisfies rank$(B^*(\lambda)) \leq i - 1$.

Proof. Because the proof of this theorem is same with that of Theorem 4.4. We only show some key difference of the proof of this theorem. Denote $\rho_2 = \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\|Y\|_F$. We have

$$\begin{aligned}
\sup_{C \in \Omega_2} \left\{ \sigma_i \left(X^T C \right) \right\} &= \sup_{\|D\|_F \leq \rho_2} \left\{ \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y + D \right) \right) \right\} \\
&= \sup_{\|D\|_F \leq \rho_2} \left\{ \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y + D \right) + X^T D \right) \right\} \\
&\leq \sup_{\|D\|_F \leq \rho_2} \left\{ \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y + D \right) \right) + \sigma_1 \left(X^T D \right) \right\} \\
&\leq \sup_{\|D\|_F \leq \rho_2} \left\{ \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \|X\|_2 \|D\|_2 \right\} \\
&\leq \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \rho_2 \|X\|_2.
\end{aligned}$$

The results of this theorem hold, if

$$\sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y + D \right) \right) + \rho_2 \|X\|_2 < 1,$$

which means

$$\sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y + D \right) \right) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) \|X\|_2 \|Y\|_F < 1. \tag{4.4}$$

To obtain the closed-form of λ , we multiply 2λ by both sides of (4.4) and get

$$2\lambda \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \|X\|_2 \|Y\|_F - \frac{\lambda}{\lambda_0} \|X\|_2 \|Y\|_F \leq 2\lambda,$$

which means

$$\begin{aligned}
\|X\|_2 \|Y\|_F &\leq 2\lambda - 2\lambda \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \frac{\lambda}{\lambda_0} \|X\|_2 \|Y\|_F \\
&= \left(2 - 2\sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \frac{1}{\lambda_0} \|X\|_2 \|Y\|_F \right) \lambda.
\end{aligned}$$

Therefore,

$$\lambda > \frac{\lambda_0 \|X\|_2 \|Y\|_F}{2\lambda_0 - 2\lambda_0 \sigma_i \left(X^T \left(C^*(\lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right) Y \right) \right) + \|X\|_2 \|Y\|_F}.$$

□

Theorem 4.7 needs that $C^*(\lambda_0)$ is known, but it is not easy to be computed for any λ_0. Next, we have a special result for $\lambda_0 = \lambda_{max}$.

Corollary 4.8. *Suppose X^TY has at least two different singular values. We define the sequence of tuning parameters $\{\lambda_i^{(2)}\}_{i=1}^r$ as*

$$\lambda_i^{(2)} = \frac{\|X^TY\|_2 \left(\|X\|_2\|Y\|_F + \sigma_i\left(X^TY\right)\right)}{2\|X^TY\|_2 - \sigma_i\left(X^TY\right) + \|X\|_2\|Y\|_F}.$$

Let $i \in \{2, \ldots, r\}$. If $\lambda \in (\lambda_i^{(2)}, \lambda_{i-1}^{(2)}]$, then

$$\sigma_j(B^*(\lambda)) = 0 \text{ holds for any } j \in \{i, \ldots, r\},$$

which leads to the solution of NMLR satisfies rank$(B^*(\lambda)) \leq i-1$.

5 Numerical studies

In this section, we introduce the popular alternating direction multiplier method (ADMM) to solve NMLR and evaluate this method on some data sets. To illustrate the efficiency of ADMM, we compare it with SLEP (Ji and Ye [14]) and CVX (Michael and Stephen [18]), which can be used to solve NMLR.

According to Theorem 3.1, the solution of NMLR can be obtained by solving (3.3). So, we give the detailed process of ADMM for solving (3.3). We first transform (3.3) as a constrained problem

$$\begin{aligned} &\min_{C,E}\left\{\frac{1}{2}\|C-Y\|_F^2 - \frac{1}{2}\|Y\|_F^2 + \delta_{\|\cdot\|_2\leq\lambda}(E)\right\} \\ &s.t. \quad X^TC - E = 0. \end{aligned} \tag{5.1}$$

Therefore, the augmented Lagrangian function is

$$\begin{aligned} L_\sigma(C, E; Z) = &\frac{1}{2}\|C-Y\|_F^2 - \frac{1}{2}\|Y\|_F^2 + \delta_{\|\cdot\|_2\leq\lambda}(E) \\ &+ \left\langle Z, X^TC - E\right\rangle + \frac{\sigma}{2}\|X^TC - E\|_F^2. \end{aligned}$$

We present the ADMM for (3.3) as follows.

Algorithm: ADMM for solving (3.3)

Step 0: Set C^0, E^0 and Z^0, let $\tau \in \left(0, \frac{1+\sqrt{5}}{2}\right)$ and $\sigma > 0$;

Step 1: Compute $C^{k+1} = \underset{C}{\operatorname{argmin}}\left\{L_\sigma\left(C, E^k; Z^k\right)\right\}$;

Step 2: Compute $E^{k+1} = \underset{E}{\operatorname{argmin}}\left\{L_\sigma\left(C^{k+1}, E; Z^k\right)\right\}$;

Step 3: Compute $Z^{k+1} = Z^k + \tau\sigma\left(X^TC^{k+1} - E^{k+1}\right)$.

Step 4: If a termination criterion is not met, go to Step 1-3.

It is easy to get the closed-form solutions of subproblems.

$$\begin{aligned}
C^{k+1} &= \underset{C}{\operatorname{argmin}} \left\{ L_\sigma \left(C, E^k; Z^k \right) \right\} \\
&= \underset{C}{\operatorname{argmin}} \left\{ \frac{1}{2}\|C - Y\|_F^2 + \left\langle X^T Z^k, C \right\rangle \right\} + \frac{\sigma}{2}\|X^T C - E^k\|_F^2 \\
&= \left(I + \sigma X X^T \right)^{-1} \left(\sigma X E^k + Y - X Z^k \right). \\
E^{k+1} &= \underset{E}{\operatorname{argmin}} \left\{ L_\sigma \left(C^{k+1}, E; Z^k \right) \right\} \\
&= \underset{E}{\operatorname{argmin}} \left\{ \delta_{\|\cdot\|_2 \le \lambda}(E) + \frac{\sigma}{2} \left\| E - X^T C^{k+1} - \frac{Z^k}{\sigma} \right\|_F^2 \right\} \\
&= \Pi_{\|\cdot\|_2 \le \lambda} \left(X^T C^{k+1} + \frac{Z^k}{\sigma} \right).
\end{aligned}$$

Here, we set the accuracy of this algorithm as $e = 10^{-4}$ and use the KKT condition of (3.3) as the terminal criterion. That is, if

$$\frac{\|(X^k)^T C^k - E^k\|_F}{1 + \|X\|_F} < e \text{ and } \frac{\|C^k - Y - X Z^k\|_F}{1 + \|Y\|_F} < e,$$

this algorithm will be stopped. Then, the algorithm returns (C^k, E^k) as the solution of (3.3), and Z^k as the solution of NMLR.

The convergence of two-blocks ADMM is well-known. For the special case (3.3), we describe its convergence result as follows (Chen et al. [5]).

Theorem 5.1. *Assume the solution set of (3.3) is nonempty. Let* $\left\{(C^k, E^k, Z^k)\right\}$ *be generated from ADMM for* $\tau \in (0, \frac{1+\sqrt{5}}{2})$. *Then the sequence* $\left\{(C^k, E^k)\right\}$ *converges to the solution of (3.3) and* $\left\{Z^k\right\}$ *converges to the solution of NMLR.*

5.1 Simulation data

Firstly, we simulate $X \in R^{n \times p}$ whose elements distribute the standard norm distribution. Secondly, we simulate the true coefficient matrix B as $B = B_1 B_2^T$, where $B_1 \in R^{p \times r}$, $B_1 \in R^{q \times r}$ and all elements of B_1 and B_2 are generated from the standard norm distribution. Hence, the rank of B is r. The error matrix W is simulated as norm distribution with mean 0 and standard variance 0.1. According to $Y = XB + W$, the response matrix is obtained. To evaluate the performance of ADMM, SLEP and CVX, we compare the computational time and test error of these three methods. To get the test error, we simulate $X_{test} \in R^{\frac{n}{2} \times p}$ and $Y_{test} = X_{test} B \in R^{\frac{n}{2} \times q}$. Then, the test error is defined as

$$\text{test error} = \frac{\|X_{test}\hat{B} - Y_{test}\|_F}{\sqrt{\frac{n}{2} \times q}},$$

where $\hat{B}$ is the solution of ADMM, SLEP or CVX. In the following tables and figures, the all reported data are the average results of 50 repetitions.

According to the results in Table 1 and Table 2, we can get the following conclusion: (i) The computational time of ADMM is slightly smaller than that of SLEP and greatly smaller than that of CVX. The computational time of CVX are larger than that of ADMM and SLEP by orders of magnitude. (ii) The test error of ADMM is smaller than that of SLEP, while the

r	time (s)			test error		
	ADMM	SLEP	CVX	ADMM	SLEP	CVX
$0.1q$	9.000e-3	1.860e-2	1.860e+1	1.647e-1	2.030e-1	2.077e-1
$0.2q$	9.300e-3	1.800e-2	1.634e+1	1.616e-1	1.827e-1	1.842e-1
$0.3q$	8.900e-3	1.780e-2	1.509e+1	1.645e-1	1.848e-1	1.770e-1
$0.4q$	1.090e-2	1.410e-2	1.635e+1	1.656e-1	2.556e-1	1.940e-1
$0.5q$	1.020e-2	1.482e-2	1.635e+1	1.685e-1	2.891e-1	1.994e-1
$0.6q$	1.000e-2	1.780e-2	1.635e+1	2.015e-1	2.294e-1	1.735e-1
$0.7q$	9.300e-3	1.800e-2	1.760e+1	1.810e-1	2.128e-1	1.760e-1
$0.8q$	1.030e-2	1.710e-2	1.512e+1	1.885e-1	2.653e-1	2.032e-1
$0.9q$	9.800e-3	1.810e-2	1.637e+1	2.087e-1	2.826e-1	1.832e-1
q	1.111e-2	1.510e-2	1.509e+1	1.864e-1	2.499e-1	1.846e-1

Table 1: Comparison between ADMM, SLEP and CVX. In this table, we set $n = 100$, $p = 80$ and $q = 10$. r means the true rank of the solution of NMLR.

r	time (s)			test error		
	ADMM	SLEP	CVX	ADMM	SLEP	CVX
$0.1q$	4.310e-2	1.184e-1	4.178e+1	1.158e-2	1.150e-1	1.462e-2
$0.2q$	5.210e-2	1.368e-1	4.154e+1	1.434e-2	1.843e-1	1.545e-2
$0.3q$	4.140e-2	1.150e-1	3.796e+1	1.632e-2	2.651e-1	1.692e-1
$0.4q$	3.560e-2	7.700e-2	4.420e+1	1.904e-2	6.958e-1	1.763e-2
$0.5q$	3.750e-2	1.232e-1	3.677e+1	2.545e-2	2.826e-1	1.891e-2
$0.6q$	3.280e-2	1.172e-1	3.605e+1	2.091e-2	2.236e-1	1.702e-2
$0.7q$	4.450e-2	1.242e-1	4.132e+1	2.881e-2	3.092e-1	2.241e-2
$0.8q$	4.210e-2	1.335e-1	3.929e+1	2.960e-2	5.179e-1	2.148e-2
$0.9q$	3.540e-2	1.408e-1	3.918e+1	2.207e-2	5.692e-1	2.234e-2
q	3.180e-2	1.221e-1	4.022e+1	2.537e-2	3.853e-1	2.472e-2

Table 2: Comparison between ADMM, SLEP and CVX. In this table, we set $n = 100$, $p = 80$ and $q = 50$. r means the true rank of the solution of NMLR.

test error of ADMM may be slightly larger than that of CVX. Therefore, we can conclude that ADMM performances better than SLEP and CVX. Because the computational time of CVX is greatly larger than that of the others, we do not compare it with other methods in the following parts. In Table 3, we increase the values of n, p and q. From the results in this table, we know that the performance of ADMM is better than that of SLEP.

In Zhou and Li [32], they simulate some data based on the signal shapes at `http://www.dabi.temple.edu/shape/MPEG7/index.html`. Following their simulation way, we set true signal shapes as B, which are reshaped as 64 times 64 matrixes. Then, we simulate X and Y as the aforementioned way. Through some experiments, we know that the larger of n, the better performance of ADMM and SLEP. Actually, these algorithms performs almost same when $n \geq 100$. To challenge the performance of ADMM and SLEP on solving NMLR, we set the sample size $n = 64$ and report the recovery of the true signal shapes under these two algorithms.

In Figure 1, because the black parts of every signal shape picture are nearly same with the true one, the main signal shapes are almost recovered with ADMM and SLEP. On the contrary, ADMM performs better than SLEP on recovering the background of these signal shape pictures. Meanwhile, the computational time of ADMM is slightly smaller than that of SLEP, from the explain of this figure. Therefore, we conclude that ADMM performs better than SLEP on recovering these signal shapes.

5.2 Real data

Now, we compare ADMM, SLEP and CVX on some real data sets. One can see the detailed information on *http://www.cad.zju.edu.cn/home/dengcai/Data/data.html*. To be understandable, we briefly introduce these real data sets.

(i) Yale: $X \in R^{165\times 1024}$, $Y \in R^{165\times 15}$.

(ii) ORL: $X \in R^{400\times 1024}$, $Y \in R^{400\times 40}$.

(iii) COIL20: $X \in R^{1440\times 1024}$, $Y \in R^{1440\times 20}$.

(iv) Isolet1-5: $X \in R^{1560\times 617}$, $Y \in R^{1560\times 26}$.

For these real data sets, the true coefficient matrix is unknown. So we evaluate the performance of ADMM and SLEP on the mean square error, which is defined as RMSE$=\frac{\|Y-X\hat{B}\|_F}{\sqrt{n*q}}$, where $\hat{B}$ is the solution of the algorithm, n and q are sizes of Y.

In Table 4, we report the performance of ADMM and SLEP on some real data sets. To be intuitively understood these results, we persent a figure about the computational time and RMSE of these algorithms. From these results, we know that ADMM performs better than SLEP on almost all real data sets. For COIL20, the computational time of ADMM is slightly larger than that of SLEP, while the RMSE of ADMM is smaller than that of SLEP.

6 Conclusion

Based on duality theory and properties of the projection operator, we give two screening rules for the nuclear norm regularized multivariate linear regression (NMLR) in high-dimensional

n, p, q	r	time (s)		test error	
		ADMM	SLEP	ADMM	SLEP
100,80,100	$0.1q$	6.610e-2	1.996e-1	1.367e-2	1.729e-1
	$0.2q$	6.920e-2	2.020e-1	1.846e-2	2.473e-1
	$0.3q$	7.450e-2	2.035e-1	1.803e-2	2.304e-1
	$0.4q$	7.170e-2	2.074e-1	1.810e-2	2.949e-1
	$0.5q$	7.480e-2	2.091e-1	1.925e-2	3.829e-1
	$0.6q$	7.560e-2	2.123e-1	1.854e-2	4.185e-1
	$0.7q$	7.330e-2	2.107e-1	1.957e-2	3.897e-1
	$0.8q$	7.690e-2	2.228e-1	2.079e-2	4.456e-1
200,80,100	$0.1q$	1.499e-1	1.541e-1	7.696e-2	7.740e-2
	$0.2q$	1.469e-1	1.478e-1	7.894e-2	7.966e-2
	$0.3q$	1.746e-1	1.424e-1	7.922e-2	8.012e-2
	$0.4q$	1.591e-1	1.604e-1	8.187e-2	8.300e-2
	$0.5q$	1.450e-1	1.633e-1	8.338e-2	8.550e-2
	$0.6q$	1.349e-1	1.693e-1	8.275e-2	8.567e-2
	$0.7q$	1.292e-1	1.508e-1	8.248e-2	9.456e-2
	$0.8q$	1.001e-1	1.864e-1	8.106e-2	8.374e-2
200,80,150	$0.1q$	1.015e-1	3.310e-1	8.066e-3	7.815e-3
	$0.2q$	1.058e-1	2.288e-1	8.610e-3	1.308e-2
	$0.3q$	1.019e-1	3.030e-1	9.101e-3	8.703e-3
	$0.4q$	1.023e-1	2.634e-1	9.346e-3	1.476e-2
	$0.5q$	1.013e-1	2.284e-1	1.010e-2	2.048e-2
	$0.6q$	1.017e-1	2.252e-1	1.103e-2	3.849e-2
	$0.7q$	1.058e-1	2.187e-1	1.085e-2	4.898e-2
	$0.8q$	1.075e-1	2.348e-1	1.099e-2	6.131e-2
200,100,200	$0.1q$	1.258e-1	4.195e-1	9.857e-2	9.936e-2
	$0.2q$	1.225e-2	4.869e-1	1.003e-1	1.008e-1
	$0.3q$	1.291e-1	4.937e-1	9.853e-2	9.870e-2
	$0.4q$	1.246e-1	3.763e-1	9.859e-2	1.180e-1
	$0.5q$	1.277e-1	5.105e-1	9.967e-2	1.014e-1
	$0.6q$	1.295e-1	5.012e-1	1.031e-1	1.044e-1
	$0.7q$	1.295e-1	4.467e-1	1.012e-1	1.116e-1
	$0.8q$	1.274e-1	5.195e-1	1.027e-1	1.043e-1

Table 3: Comparison between ADMM and SLEP.

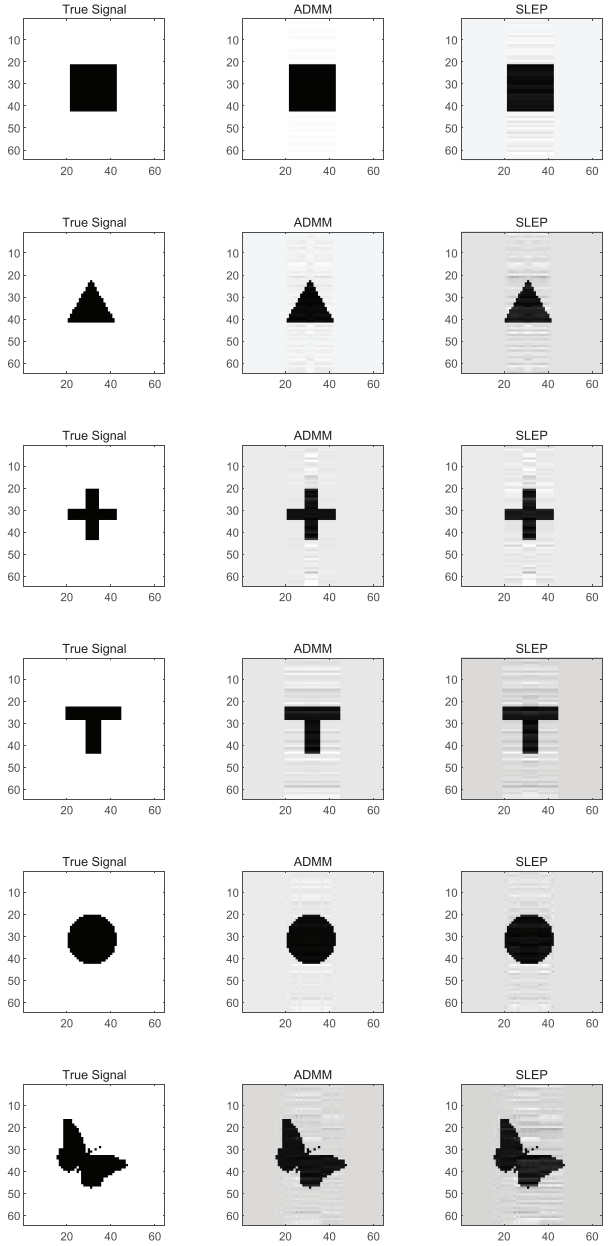

Figure 1: The comparison of ADMM and SLEP on some signal shapes. The computational time of ADMM on recovery these signa shapes are 0.026s, 0.037s, 0.033s, 0.027s, 0.027s and 0.032s, respectively. The computational time of SLEP on recovery these signa shapes are 0.069s, 0.080s, 0.080s, 0.067s, 0.087s and 0.081s, respectively.

data sets	time (s)		RMSE	
	ADMM	SLEP	ADMM	SLEP
Yale	2.810e-1	3.820e-1	9.550e-2	1.106e-1
ORL	5.730e-1	2.037e+1	6.370e-2	1.175e-1
COIL20	2.333e+1	2.141e+1	2.630e-2	1.096e-1
Isolet1	1.428e+1	1.986e+1	8.740e-2	1.096e-1
Isolet2	1.502e+1	2.009e+1	8.810e-2	1.110e-1
Isolet3	1.445e+1	2.004e+1	9.380e-2	1.158e-1
Isolet4	1.373e+1	2.014e+1	9.511e-2	1.165e-1
Isolet5	1.874e+1	1.966e+1	9.370e-2	1.168e-1

Table 4: Comparison of ADMM and SLEP on real data sets.

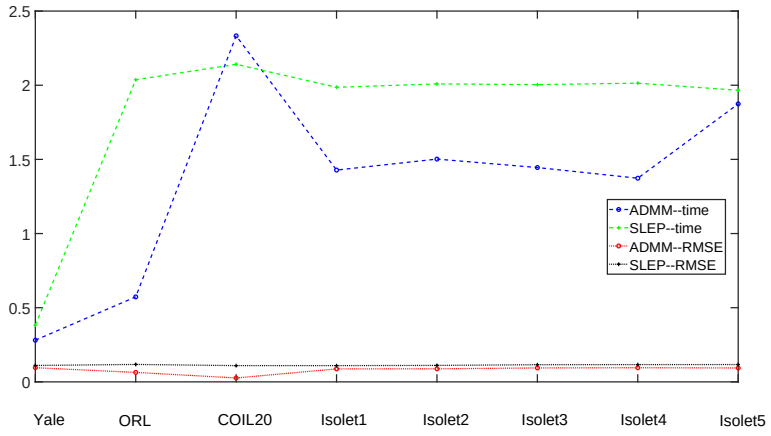

Figure 2: The comparison of ADMM and SLEP on real data sets.

setting. For any tuning parameter, these rules identify the inactive singular values and estimate the maximal rank of the solution of NMLR. Furthermore, we propose ADMM to solve NMLR and evaluate the performance of this method on some numerical experiments.

Acknowledgements

We sincerely thank the referees as well as the associate editor for their constructive comments which have significantly improved the quality of the paper.

References

[1] H. Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, *J. Environ. Health. Sci.*, 12 (2014) 1–40.

[2] A. Agarwal, S. Negahban and M. J. Wainwright, Fast global convergence of gradient methods for high-dimensional statistical recovery, *Ann. Statist.* 40 (2012) 2452–2482.

[3] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1984.

[4] L. Bottou, E. F. Curtis and J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.* 60 (2018) 223–311.

[5] L. Chen, D. Sun and K. Toh, A note on the convergence of ADMM for linearly constrained convex optimization problems, *Comput. Optim. Appl.* 66 (2017) 327–343.

[6] S. Chen, D. Donoho and M. Saunders, Atomic decomposition for basis pursuit, *SIAM J. Sci. Comput.* 20 (1998) 33–61.

[7] D. R. Cox, The regression analysis of binary sequences, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 20 (2008) 215–242.

[8] E. L. Ghaoui, V. Viallon and T. Rabbani, Safe feature elimination in sparse supervised learning, *Pacific J. Optim.* 8 (2012) 667–698.

[9] J. Fan and J. Lv, Sure independence screening for ultrahigh dimensional feasure space (with discussion), *J. R. Stat. Soc. Ser. B Stat. Methodol.* 70 (2008) 849–911.

[10] Y. Fan and C. Y. Tang, Tuning parameter selection in high dimensional penalized likelihood, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 75 (2013) 531–552.

[11] M. Fazek, T. K. Pong, D. Sun and P. Tseng, Hankel matrix rank minization with applications to system identification and realization, *SIAM J. Matrix Anal. Appl.* 34 (2013) 946–977.

[12] K. H. Francis, I. W. David and D. F. Scott, Tuning parameter selection for the adaptive lasso using ERIC, *J. Amer. Statist. Assoc.* 110 (2015) 262–269.

[13] L. Kong, J. Sun, J. Tao and N. Xiu, Sparse recovery on Euclidean jordan algebras, *Linear Algebra Appl.* 465 (2015) 65–87.

[14] S. Ji and J. Ye, An accelerated gradient method for trace norm minimization. in: *ICML*, 2009, pp. 457–464.

[15] Z. B. Kuang, S. N. Geng and D. Page, A screening rule for l_1-regularized ising model estimation, in: *NIPS*, vol. 30, 2017, pp. 720–731.

[16] S. Lee, N. Gornitz, E. P. Xing, D. Heckerman and C. Lippert, Ensembles of Lasso screening rules, *IEEE Trans. Pattern Anal.* 40 (2018) 2841–2852.

[17] S. Ma, R. Li and C. L. Tsai, Variable screening via quantile partial correlation, *J. Amer. Statist. Assoc.* 112 (2017) 650–663.

[18] G. Michael and B. Stephen, CVX: Matlab software for disciplined convex programming, version 2.0 beta, (2013) http://cvxr.com/cvx.

[19] E. Ndiaye, O. Fercoq, A. Gramfort and J. Salmon, Gap safe screening rules for sparsity enforcing penalties, *J. Mach. Learn. Res.* 18 (2017) 1–33.

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

[21] R. Noorossana, M. Eyvazian, A. Amiri and M. A. Mahmoud, Statistical monitoring of multivariate multiple linear regression profiles in phase I with calibration application, *Qual. Reliab. Eng. Int.* 26 (2010) 291–303.

[22] X. Pan and Y. Xu, A safe reinforced feature screening strategy for lasso based on feasible solutions, *Inform. Sciences* 477 (2019) 132–147.

[23] J. Peng, L. Zhu, A. Bergamaschi, W. Han, D. Y. Noh, J. R. Pollack and P. Wang, Regularized multivariate regression for idetifying master predictors with application to integrative genomics study of breast cancer, *Ann. Appl. Stat.*, 4 (2012) 53–77.

[24] R. T. Rockafellar, *Convex Analysis*, Princeton University, Princeton, 1970.

[25] A. H. Roger, *Matrix Analysis,* 2nd edn, Cambridge University, Cambridge, 2013.

[26] R. Tibshirani, J. Bien, T. Hastie, N. Simon, J. Taylor and R. J. Tibshirani, Strong rules for discarding predictors in lasso-type problems, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74 (2012) 1–22.

[27] R. Tibshirani Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58 (1996) 267–288.

[28] J. Wang, P. Wonka and J. Ye, Lasso screening rules via dual polytope projection, *J. Mach. Learn. Res.* 16 (2015) 1063–1101.

[29] G. A. Watson, Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.* 170 (1992) 33–45.

[30] Z. J. Xiang, Y. Wang and J. P. Ramadge, Screening tests for Lasso problems, *IEEE Trans. Pattern. Anal.* 5 (2017) 1008–1027.

[31] M. Yuan, A. Ekici, Z. Lu and R. Monteiro, Dimension reduction and coefficient estimation in multivariate linear regression, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69 (2007) 329–346.

[32] H. Zhou and L. X. Li. Regularized matrix regression, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 76 (2014) 463–483.

Manuscript received 23 October 2020
revised 2 January 2021
accepted for publication 6 February 2021

Pan Shang
Department of Applied Mathematics, Beijing Jiaotong University
Address: N0.3 Shangyuancun, Haidian District, Beijing, 100044, P. R. China
E-mail address: 18118019@bjtu.edu.cn

Lingchen Kong
Department of Applied Mathematics, Beijing Jiaotong University
Address: N0.3 Shangyuancun, Haidian District, Beijing, 100044, P. R. China
E-mail address: konglchen@126.com