Regularization Parameter Selection for the Low Rank Matrix Recovery

Pan Shang · Lingchen Kong

Received: 16 September 2019 / Accepted: 24 March 2021 / Published online: 9 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
A popular approach to recover low rank matrices is the nuclear norm regularized minimization (NRM) for which the selection of the regularization parameter is inevitable. In this paper, we build up a novel rule to choose the regularization parameter for NRM, with the help of the duality theory. Our result provides a safe set for the regularization parameter when the rank of the solution has an upper bound. Furthermore, we apply this idea to NRM with quadratic and Huber functions, and establish simple formulae for the regularization parameters. Finally, we report numerical results on some signal shapes by embedding our rule into the cross validation, which state that our rule can reduce the computational time for the selection of the regularization parameter. To the best of our knowledge, this is the first attempt to select the regularization parameter for the low rank matrix recovery.

Keywords Regularization parameter selection rule · Low rank matrix recovery · Nuclear norm regularized minimization · Duality theory

Mathematics Subject Classification 90C46 · 65K05 · 90C06 · 90C25 · 62J99

1 Introduction

The low rank matrix recovery (LMR) arises in tremendous applications, such as machine learning [4], signal processing [5], system identification [18], biomedical imaging [32] and so on. Since it is an NP-hard problem, an alternative method to
recover low rank matrices is the nuclear norm minimization. In fact, the nuclear norm minimization can yield the exact solution of LMR under some assumptions, such as the restricted isometry property (RIP, e.g., Recht et al. [22], Cai and Zhang [2]) and the s-goodness condition (e.g., Kong et al. [15]). In statistics, a popular way to solve the nuclear norm minimization is the nuclear norm regularized minimization (NRM). See, e.g., Koltchinskii et al. [14], Mark and Justin [5], Bottou and Nocedal [1], Lu et al. [19], Negahban and Wainwright [21], Rohde and Tsybakov [24], Yuan et al. [31], Zhou and Li [33].

Regularization parameter selection plays an essential role for solving the regularization models. In practice, the cross validation is a common method to choose this parameter. However, its computational cost is usually expensive. For LASSO (Chen et al. [3] and Tibshirani [27]), some screening rules have been proposed to eliminate inactive features under different regularization parameters. Therefore, screening rules can be used to accelerate the cross validation and select the regularization parameter. See, e.g., Fan and Lv [7], Ghaoui et al. [10], Tibshirani et al. [26], Wang et al. [28], Ndiaye et al. [20], Kuang et al. [16], Xiang et al. [30], Lee et al. [17]. Existing screening rules can be roughly divided into two categories: safe rules and heuristic rules. The safe rule means that the discarded features are guaranteed to be inactive, therefore there are a lot of research about this rule. For instance, Ghaoui et al. [10] constructed SAFE rules to eliminate predictors and these rules never remove active predictors. Ndiaye et al. [20] built up statics and dynamic gap safe screening rules which are based on the gap between feasible points of LASSO and its dual problem. Meanwhile, there are some heuristic rules, which can not guarantee that eliminated features are inactive. For example, Tibshirani et al. [26] proposed strong rules for discarding inactive predictors under the unit slope bound assumption. Although strong rules are heuristic, they screen out far more predictors than SAFE rules in practice and can be more effective by checking Karush-Kuhn-Tucker (KKT) conditions. However, to the best of our knowledge, there are no regularization parameter selection results for NRM, which help to improve the efficiency of the cross validation. Note that NRM includes the sparse vector selection (compress sensing) as a special case. Thus, NRM reduces into LASSO type problems in some special cases. One natural question is: can we establish the regularization parameter selection result for NRM?

In this paper, we give an affirmative answer. In order to do so, we present the dual form of NRM and the strong duality theorem. Then, we propose a regularization parameter selection rule for NRM based on its dual solution. If the rank of the solution has an upper bound, this result provides a safe set of the regularization parameter. This is the primal result that can help select the regularization parameter, but it may be difficult to get the dual solution. By analyzing the loss function of NRM and the duality gap, we obtain an implementable rule that depends on feasible points of NRM and its dual problem. In particular, this idea is applied to the nuclear norm regularized quadratic minimization (Q-NRM) and the nuclear norm regularized Huber minimization (H-NRM). For every problem, the regularization parameter selection rule presents a sequence of closed-form parameters and the maximal rank of the solution under these parameters. According to these results, the cross validation can be implemented on the provided set to select the regularization parameter. For the purpose of enlightening the regularization parameter selection rule, we consider some
signal shapes that were analyzed in Zhou and Li [33]. Numerical results show that our rule can reduce the computational time for the selection of the regularization parameter.

The rest of the paper is organized as follows. We review some related models and build up the duality theory of NRM in Sect. 2. In Sect. 3, we introduce the regularization parameter selection rule for NRM. In Sect. 4, we apply the regularization parameter selection rule to Q-NRM and H-NRM. In Sect. 5, we present some numerical results of the regularization parameter selection rule. Some conclusions are given in Sect. 6.

Notations: For any vector \(x \in \mathbb{R}^n\), the 2-norm \(\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}\).

For any matrix \(M \in \mathbb{R}^{p \times q}\), suppose \(M\) has a singular value decomposition with nonincreasing singular values \(\sigma_1(M) \geq \cdots \geq \sigma_r(M) \geq 0\), where \(r = \min\{p, q\}\). There are some norms based on singular values of \(M\). The Frobenius norm \(\|M\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^q M_{ij}^2 = \sum_{i=1}^r \sigma_i^2(M) + \cdots + \sigma_r^2(M)}\) is defined as \(\|\cdot\|_F\). 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., \(\|M\|_2 = \sigma_1(M)\).

2 Preliminaries

In this section, we review the low rank matrix recovery (LMR) and its convex relaxation, which replace the rank function by the nuclear norm. Moreover, we analyze the nuclear norm regularized minimization (NRM) and build up its duality theory.

LMR problem finds a matrix with the minimum rank from some linear constraints, that is

\[
\min_{B \in \mathbb{R}^{p \times q}} \text{rank}(B)
\]

\[
\text{s.t. } Y = A(B) + \epsilon, \quad \|\epsilon\|_2 \leq \delta,
\]

where \(A(B) = ((X_1, B), \ldots, (X_n, B))^T\), \(Y = (y_1, y_2, \ldots, y_n)^T\) with given \((X_i, y_i) \in \mathbb{R}^{p \times q} \times \mathbb{R} (i \in \{1, 2, \ldots, n\})\), \(\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T\) is the noise vector and \(\delta \geq 0\) is a constant. If \(\delta = 0\), the model is the noiseless LMR. Otherwise, it is the noisy LMR. It is an NP-hard problem because rank\((B)\) is noncontinuous and nonconvex. However, LMR has many important applications in statistics, machine learning and so on. See, e.g., Chen et al. [4], Davenport and Romberg [5], Fazel et al. [9], Liu and Vandenberghe [18], Zhao et al. [32].

An alternative method to recover low rank matrices is to solve the convex relaxation of LMR, which is the following nuclear norm minimization (e.g., Fazel [8])

\[
\min_{B \in \mathbb{R}^{p \times q}} \|B\|_*
\]

\[
\text{s.t. } Y = A(B) + \epsilon, \quad \|\epsilon\|_2 \leq \delta.
\]

Note that there are a lot of research on the theoretical guarantee that this relaxation problem yields the exact solution of LMR under some conditions, such as the restricted
isometry property (see, e.g., Recht et al. [22], Cai and Zhang [2]) and the s-goodness condition (e.g., Kong et al. [14]). In statistics, a popular method is NRM, which is the unconstraint optimization problem of the nuclear norm minimization. One common choice is the nuclear norm regularized quadratic minimization (Q-NRM), see (7) in Sect. 4. When the noise $\epsilon$ has zero mean and constant variance, Q-NRM performs well in recovering the low rank matrix. However, it is not efficient when the noise has heavy tails or outliers. In fact, the distribution of noise is unknown in practice. In this sense, another recently attractive model is the nuclear norm regularized Huber minimization (H-NRM). See, e.g., Huber [13], Sun [25], Elsener and Geer [6]. See (10) in Sect. 4.

How to choose the regularization parameter for NRM is an essential question. Many works on Q-NRM and H-NRM often use the cross validation to select this parameter. To the best of our knowledge, there are no theoretical results on the selection of the regularization parameter for NRM. In order to establish this kind of result, we introduce NRM as follows.

$$\min_{B \in \mathbb{R}^{p \times q}} \left\{ F_\lambda(B) = \sum_{i=1}^{n} f_i (y_i - \langle X_i, B \rangle) + \lambda \|B\|_* \right\},$$

where $f_i : \mathbb{R} \mapsto \mathbb{R}$ is a proper, closed and convex function with $\frac{1}{\alpha}$-Lipschitz continuous gradient ($\alpha > 0$). Note that this problem includes Q-NRM and H-NRM as special cases. In order to address that the solution of NRM depends on the regularization parameter $\lambda$, we denote it as $B^*(\lambda)$. The assumptions on $\{f_i\}_{i=1}^{n}$ are commonly used in the literature. See, e.g., Ndiaye et al. [20], where they consider a group-decomposable regularizer.

The duality theory plays an important role in obtaining the results for the selection of the regularization parameter. Thus, we consider the dual problem of NRM. By introducing new variables $t_i = y_i - \langle X_i, B \rangle$, $i \in \{1, 2, \ldots, n\}$, we rewrite NRM as

$$\min_{B \in \mathbb{R}^{p \times q}, t \in \mathbb{R}^n} \left\{ F_\lambda(B, t) = \sum_{i=1}^{n} f_i (t_i) + \lambda \|B\|_* \right\}$$

s.t. $y_i - \langle X_i, B \rangle - t_i = 0$, $i \in \{1, 2, \ldots, n\}$,

where $t = (t_1, t_2, \ldots, t_n)^T$. Then, the Lagrangian function of (2) is

$$L(B, t; \theta) = \sum_{i=1}^{n} f_i (t_i) + \lambda \|B\|_* + \sum_{i=1}^{n} \theta_i \cdot (y_i - \langle X_i, B \rangle - t_i),$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ with $\theta_i \in \mathbb{R}$, $\forall i \in \{1, 2, \ldots, n\}$ being the Lagrangian multiplier of (2). By direct computation, we obtain that
\[
\begin{align*}
\min_{B \in \mathbb{R}^{p \times q}, t \in \mathbb{R}^n} & \quad L(B, t; \theta) \\
& = \min_{B \in \mathbb{R}^{p \times q}} \left\{ \lambda \|B\|_* - \left( \sum_{i=1}^n \theta_i X_i, B \right) \right\} + \min_{t \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \left( f_i(t_i) - \theta_i t_i \right) \right\} + \langle y, \theta \rangle \\
& = -\max_{B \in \mathbb{R}^{p \times q}} \left\{ \left( \sum_{i=1}^n \theta_i X_i, B \right) - \lambda \|B\|_* \right\} - \sum_{i=1}^n \max_{t_i \in \mathbb{R}} \left\{ \theta_i t_i - f_i(t_i) \right\} + \langle y, \theta \rangle.
\end{align*}
\]

Based on the definition of the conjugate function (Rockafellar [23]), we know that
\[
\max_{B \in \mathbb{R}^{p \times q}} \left\{ \left( \sum_{i=1}^n \theta_i X_i, B \right) - \lambda \|B\|_* \right\} = \begin{cases} 0, & \|\sum_{i=1}^n \theta_i X_i\|_2 \leq \lambda \\ +\infty, & \text{otherwise} \end{cases}
\]
and \( \max_{t_i \in \mathbb{R}} \left\{ \theta_i t_i - f_i(t_i) \right\} = f_i^*(\theta_i) \). Thus,
\[
\min_{B \in \mathbb{R}^{p \times q}, t \in \mathbb{R}^n} L(B, t; \theta) = \begin{cases} -\sum_{i=1}^n f_i^*(\theta_i) + \langle y, \theta \rangle, & \|\sum_{i=1}^n \theta_i X_i\|_2 \leq \lambda \\ -\infty, & \text{otherwise} \end{cases}
\]

Therefore, the dual problem of (2) is
\[
\begin{align*}
\max_{\theta \in \mathbb{R}^n} & \quad G_\lambda(\theta) = \langle y, \theta \rangle - \sum_{i=1}^n f_i^*(\theta_i) \\
\text{s.t.} & \quad \|\sum_{i=1}^n \theta_i X_i\|_2 \leq \lambda.
\end{align*}
\]

Denote the solution of (3) as \( \theta^*(\lambda) \). The Karush-Kuhn-Tucker (KKT) system of (2) is
\[
\begin{align*}
\sum_{i=1}^n \theta_i X_i & \in \lambda \partial \|B\|_*, \\
y_i - \langle X_i, B \rangle - t_i &= 0, \\
\theta_i &= \nabla f_i(t_i), \quad i = 1, 2, \ldots, n.
\end{align*}
\]

If a triple \( (B(\lambda), t(\lambda), \theta(\lambda)) \) satisfies the KKT system, it is called the KKT point of (2). Note that \( (0, y) \) is a feasible point of (2). Slater constraint qualification holds on this problem, which leads to the following duality theorem.

**Theorem 2.1** (Strong duality theorem) *If the solution of (2) exists, then there is a KKT point \( (B^*(\lambda), t^*(\lambda), \theta^*(\lambda)) \) such that the optimal values of (2) and (3) are equal, i.e.,
\[
F_\lambda(B^*(\lambda), t^*(\lambda)) = G_\lambda(\theta^*(\lambda)).
\]
Here, \( B^*(\lambda) \) is the solution of NRM and \( \theta^*(\lambda) \) is the solution of (3).

### 3 Regularization Parameter Selection Rule

With the help of the duality theory in Sect. 2, we will present the general regularization parameter selection rule for the nuclear norm regularized minimization (NRM), which clarifies the regularization parameter can be selected by feasible points of NRM and its dual problem.

It is clear that the solution of NRM is zero when \( \lambda \) is sufficiently large. In the following lemma, we show the connection between the zero solution of NRM and \( \lambda \). This result is inspired by Ndiaye et al. [20, Proposition 4], where they consider the group-decomposable regularizer. However, the nuclear norm in NRM is not group-decomposable.

**Lemma 3.1** Let \( \theta_{\text{max}} = \arg\max_{\theta \in \mathbb{R}^n} \{ \langle y, \theta \rangle - \sum_{i=1}^n f_i^*(\theta_i) \} \) and

\[
\lambda_{\text{max}} = \left\| \sum_{i=1}^n \theta_{\text{max}}^i X_i \right\|_2.
\]

The following statements hold.

(i) If \( B^*(\lambda) = 0 \), then \( \lambda \geq \lambda_{\text{max}} \).

(ii) If \( \lambda > \lambda_{\text{max}} \), then \( B^*(\lambda) = 0 \).

**Proof** (i) Since \( B^*(\lambda) = 0 \), by Theorem 2.1,

\[
F_{\lambda}(0, y) = G_{\lambda}(\theta^*(\lambda)).
\]

It is true that \( G_{\lambda}(\theta^*(\lambda)) \leq \max_{\theta \in \mathbb{R}^n} \langle y, \theta \rangle - \sum_{i=1}^n f_i^*(\theta_i) \). Since \( \{ f_i \}_{i=1}^n \) are closed and convex,

\[
G(\theta_{\text{max}}) = \sum_{i=1}^n f_i(y_i) = F_{\lambda}(0, y).
\]

Combining these results, we know that

\[
F_{\lambda}(0, y) = G_{\lambda}(\theta^*(\lambda)) \leq G(\theta_{\text{max}}) = F_{\lambda}(0, y).
\]

Because \( f_i \) is differentiable and its gradient is \( \frac{1}{\alpha} \)-Lipschitz continuous, its conjugate function \( f_i^* \) is \( \alpha \)-strongly convex (Hiriart-Urruty and Lemaréchal [11]). Then, the function \( G_{\lambda}(\theta) \) is \( \alpha \alpha \)-strongly concave and (3) has a unique solution. Therefore, \( \theta_{\text{max}} \) is a solution of (3) and \( \left\| \sum_{i=1}^n \theta_{\text{max}}^i X_i \right\|_2 \leq \lambda \). Hence, the desired result follows.
(ii) If $\lambda > \lambda_{\text{max}}$, we can get $\| \sum_{i=1}^{n} \theta_i \max X_i \|_2 \leq \lambda$ and $\theta^*(\lambda) = \theta^\max$. By Theorem 2.1, we know that $G_{\lambda}(\theta^\max) = F_{\lambda}(B^*(\lambda), t^*(\lambda))$. Because $(0, y)$ is a feasible point of (2), it is clear that $F_{\lambda}(B^*(\lambda), t^*(\lambda)) \leq F_{\lambda}(0, y)$. According to the convexity of $f_i$, $F_{\lambda}(0, y) = G_{\lambda}(\theta^\max)$. Therefore, we know that

$$G_{\lambda}(\theta^\max) = F_{\lambda}(B^*(\lambda), t^*(\lambda)) \leq F_{\lambda}(0, y) = G_{\lambda}(\theta^\max).$$

Clearly, zero is a solution of (1) when $\lambda > \lambda_{\text{max}}$. Moreover, $\sum_{i=1}^{n} \theta_i \max X_i \in \lambda \partial \| B^*(\lambda) \|_*$. Thus, $B^*(\lambda) = 0$ is the unique solution.

From Lemma 3.1, without loss of generality, we may focus on the regularization parameter $\lambda$ in $(0, \lambda_{\text{max}}]$ with $\lambda_{\text{max}} > 0$. Because NRM tends to obtain the low rank solution, we choose 0 as its solution when $\lambda = \lambda_{\text{max}}$. We next show a basic result which provides a relationship between the regularization parameter $\lambda$ and the maximal rank of the solution of NRM.

**Theorem 3.1** For any $k \in \{1, 2, \ldots, r\}$ with $r = \min\{p, q\}$, if

$$\lambda > \sigma_k \left( \sum_{i=1}^{n} \theta_i^*(\lambda) X_i \right),$$

then $\sigma_k(B^*(\lambda)) = 0$. This leads to $\text{rank}(B^*(\lambda)) \leq k - 1$.

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

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

where

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

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

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

By the KKT system, we have

$$\sum_{i=1}^{n} \theta_i^*(\lambda) X_i \in \lambda \partial \| B^*(\lambda) \|_*.$$
Combining this result with the subdifferential of $B^*(\lambda)$, we get

$$\sigma_k \left( \sum_{i=1}^{n} \theta_i^*(\lambda)X_i \right) \in \begin{cases} [\lambda], & \sigma_k(B^*(\lambda)) > 0 \\ [0, \lambda], & \sigma_k(B^*(\lambda)) = 0 \end{cases}, \quad k \in \{1, 2, \ldots, r\}.$$ 

Therefore, if $\lambda > \sigma_k \left( \sum_{i=1}^{n} \theta_i^*(\lambda)X_i \right)$, we must have $\sigma_k(B^*(\lambda)) = 0$, which leads to $\text{rank}(B^*(\lambda)) \leq k - 1$.

This theorem estimates the maximal rank of the solution of NRM under different regularization parameters. If the rank of the solution of NRM has an upper bound $k - 1$, this theorem provides a safe interval

$$(\sigma_k(\sum_{i=1}^{n} \theta_i^*(\lambda)X_i), \lambda_{\max}]$$

that is smaller than $(0, \lambda_{\max}]$. Therefore, embedding this rule into the cross validation may reduce the computational time for the selection of the regularization parameter. Notice that we first need to get the solution of (3) when we apply Theorem 3.1 to choose the regularization parameter for NRM. However, the dual solution may be hard to be obtained. Usually, we can easily obtain a feasible set which contains the dual solution. Denote the duality gap of (2) and (3) as $\text{Gap}(\lambda) = F_\lambda(B(\lambda), t(\lambda)) - G_\lambda(\theta(\lambda))$. Ndiaye et al. [20, Theorem 6] gives a feasible set based on assumptions of $f_i$ and the duality gap, it holds on our problem. We review this in the next lemma.

**Lemma 3.2** For any feasible points $(B(\lambda), t(\lambda))$ of (2) and $\theta(\lambda)$ of (3), it holds that

$$\|\theta(\lambda) - \theta^*(\lambda)\| \leq \sqrt{2\text{Gap}(\lambda)/n\alpha}. \quad (5)$$

Lemma 3.2 means that the dual solution $\theta^*(\lambda)$ is contained in the set

$$B = \left\{ \theta \bigg| \|\theta - \theta(\lambda)\| \leq \sqrt{2\text{Gap}(\lambda)/n\alpha} \right\}.$$ 

According to Theorem 3.1, if

$$\max_{\theta \in B} \sigma_k \left( \sum_{i=1}^{n} \theta_iX_i \right) < \lambda, \quad (6)$$

then $\sigma_k(B^*(\lambda)) = 0$. Combining Theorem 3.1 and Lemma 3.2, we obtain the following regularization parameter selection rule.
Theorem 3.2 Let \((B(\lambda), t(\lambda))\) and \(\theta(\lambda)\) be any feasible points of (2) and (3), respectively. For any \(k \in \{1, 2, \ldots, r\}\) with \(r = \min\{p, q\}\), if
\[
\lambda > \sigma_k \left( \sum_{i=1}^{n} \theta_i(\lambda) X_i \right) + \sqrt{\frac{2\text{Gap}(\lambda)}{\alpha} \cdot \frac{\sum_{i=1}^{n} \|X_i\|_2^2}{n}},
\]
then \(\sigma_k(B^*(\lambda)) = 0\). This leads to \(\text{rank}(B^*(\lambda)) \leq k - 1\).

Proof For any regularization parameter \(\lambda > 0\) and any feasible point \(\theta(\lambda)\) of (3), Lemma 3.2 shows that the dual solution \(\theta^*(\lambda)\) is contained in the set
\[
B = \left\{ \theta \left| \|\theta - \theta(\lambda)\|_2 \leq \gamma \right. \right\},
\]
where \(\gamma = \sqrt{\frac{2\text{Gap}(\lambda)}{n\alpha}}\). Setting \(\eta = \theta - \theta(\lambda)\), we have
\[
\max_{\theta \in B} \sigma_k \left( \sum_{i=1}^{n} \theta_i X_i \right) = \max_{\|\eta\|_2 \leq \gamma} \sigma_k \left( \sum_{i=1}^{n} (\theta_i(\lambda) + \eta_i) X_i \right).
\]

Following the singular value inequalities in Horn and Johnson [12] (see, page 454), we get
\[
\sigma_k \left( \sum_{i=1}^{n} (\theta_i(\lambda) + \eta_i) X_i \right) \leq \sigma_k \left( \sum_{i=1}^{n} \theta_i(\lambda) X_i \right) + \sigma_1 \left( \sum_{i=1}^{n} \eta_i X_i \right).
\]

Then
\[
\max_{\|\eta\|_2 \leq \gamma} \sigma_k \left( \sum_{i=1}^{n} (\theta_i(\lambda) + \eta_i) X_i \right) \leq \sigma_k \left( \sum_{i=1}^{n} \theta_i(\lambda) X_i \right) + \max_{\|\eta\|_2 \leq \gamma} \sigma_1 \left( \sum_{i=1}^{n} \eta_i X_i \right)
\]
\[
= \sigma_k \left( \sum_{i=1}^{n} \theta_i(\lambda) X_i \right) + \max_{\|\eta\|_2 \leq \gamma} \left\| \sum_{i=1}^{n} \eta_i X_i \right\|_2.
\]

According to the triangle inequality of the spectral norm \(\| \cdot \|_2\), we have
\[
\left\| \sum_{i=1}^{n} \eta_i X_i \right\|_2 \leq \sum_{i=1}^{n} |\eta_i| \cdot \|X_i\|_2 = \langle |\eta|, x \rangle,
\]
where $|\eta| = (|\eta_1|, |\eta_2|, \ldots, |\eta_n|)^T$ and $x = (\|X_1\|_2, \|X_2\|_2, \ldots, \|X_n\|_2)^T$. By the Cauchy-Buniakowsky-Schwarz inequality,

$$
\langle |\eta|, x \rangle \leq \||\eta||\|_2 \|x\|_2 = \|\eta\|_2 \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2}.
$$

Therefore,

$$
\max_{\|\eta\|_2 \leq \gamma} \sum_{i=1}^{n} \eta_i X_i \|_2 \leq \max_{\|\eta\|_2 \leq \gamma} \|\eta\|_2 \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2} = \gamma \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2}.
$$

So, we have that

$$
\max_{\|\eta\|_2 \leq \gamma} \sigma_k \left( \sum_{i=1}^{n} (\theta_i(\lambda) + \eta_i) X_i \right) \leq \sigma_k \left( \sum_{i=1}^{n} \theta_i(\lambda) X_i \right) + \gamma \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2}.
$$

The desired result follows from Theorem 3.1 and (6). $\square$

If the rank of the solution of NRM has an upper bound, Theorem 3.2 implies that the safe interval of $\lambda$ can be decided by the feasible points of (2) and (3). Comparing to Theorem 3.1, this result is easier to be implemented. It is clear that Theorem 3.2 reduces to Theorem 3.1, when $\text{Gap}(\lambda) = 0$ and $B = \{\theta^\ast(\lambda)\}$.

Actually, one can choose any feasible points of (2) and (3) to obtain $\text{Gap}(\lambda)$. For instance, we give the following approach to obtain primal and dual feasible points. Let $\theta(\lambda) = \frac{\lambda \theta_{\lambda_{\max}}}{\lambda_{\max}}$ with $\lambda \in (0, \lambda_{max})$. Clearly, $\theta(\lambda)$ is a feasible point of (3). Moreover, it is easy to check that

$$
B(\lambda) = \frac{\sum_{i=1}^{n} \theta_i(\lambda) X_i}{\lambda}_{\lambda_{\max}} = \frac{\sum_{i=1}^{n} \theta_i^\max X_i}{\lambda_{\max}}
$$

and $t_i(\lambda) = y_i - \langle X_i, B(\lambda) \rangle (i \in \{1, 2, \ldots, n\})$ are feasible points of (2). Clearly, these feasible points relate to $\theta^\max$.

Based on the definition of $\theta^\max$ in Lemma 3.1, the above feasible points depend on the specific form of $f_i$. In the next section, we give two specific forms of $f_i$, which are the quadratic function and the Huber function. For the nuclear norm regularized quadratic minimization (Q-NRM) and the nuclear norm regularized Huber minimization (H-NRM), we show efficient results for the selection of the regularization parameter.
4 Two Specific Applications: Q-NRM and H-NRM

This section illustrates applications of the regularization parameter selection rule on Q-NRM and H-NRM.

4.1 Regularization Parameter Selection Rule for Q-NRM

It is well known that Q-NRM is a popular method to recover the low rank matrix. Q-NRM is given as

\[
\min_{B \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle X_i, B \rangle)^2 + \lambda \| B \|_* \right\}.
\] (7)

We denote the solution of Q-NRM by \( B^{\ast(\text{ls})}(\lambda) \). Here, \( f_i(\mu) = \frac{1}{2} \mu^2 \). Clearly, it is a differentiable function with 1-Lipschitz continuous gradient. Interestingly, the dual form of (7) is a projection problem on a compact and convex area, which is given as

\[
\max_{\theta \in \mathbb{R}^n} \left\{ \langle y, \theta \rangle - \frac{1}{2} \| \theta \|_2^2 \right\}
\text{s.t. } \left\| \sum_{i=1}^{n} \theta_i X_i \right\|_2 \leq \lambda.
\] (8)

Let \( \theta^{\max(\text{ls})} = \arg \max_{\theta \in \mathbb{R}^n} \left\{ \langle y, \theta \rangle - \frac{1}{2} \| \theta \|_2^2 \right\} = y \) and \( \lambda^{(\text{ls})}_{\max} = \left\| \sum_{i=1}^{n} y_i X_i \right\|_2 \). From Lemma 3.1, \( \lambda^{(\text{ls})}_{\max} \) is the lower bound of the regularization parameter such that the solution of Q-NRM is zero, and \( \theta^{\max(\text{ls})} \) the solution of (8) with \( \lambda = \lambda^{(\text{ls})}_{\max} \). Then, for any \( \lambda \in (0, \lambda^{(\text{ls})}_{\max}) \), we can easily give the feasible points of (7) and (8) as

\[
B_0^{(\text{ls})}(\lambda) = \frac{\sum_{i=1}^{n} y_i X_i}{\lambda^{(\text{ls})}_{\max}} \quad \text{and} \quad \theta_0^{(\text{ls})}(\lambda) = \frac{\lambda y}{\lambda^{(\text{ls})}_{\max}}.
\]

Therefore, the dual solution is contained in the set

\[
\theta \bigg| \| \theta - \theta_0^{(\text{ls})}(\lambda) \|_2 \leq \sqrt{\frac{2\text{Gap}^{(\text{ls})}(\lambda)}{n}},
\]

where \( \text{Gap}^{(\text{ls})}(\lambda) \) is

\[
\frac{1}{2} \sum_{i=1}^{n} \left( y_i - \frac{\langle X_i, \sum_{i=1}^{n} y_i X_i \rangle}{\lambda^{(\text{ls})}_{\max}} \right)^2 + \frac{\lambda}{\lambda^{(\text{ls})}_{\max}} \left\| \sum_{i=1}^{n} y_i X_i \right\|_* - \left( \frac{\lambda}{\lambda^{(\text{ls})}_{\max}} - \frac{1}{2} \right) \| y \|_2^2.
\]
In order to show the closed form of the regularization parameters, we define some new notations.

\[
\lambda_0^{(ls)} = \lambda_{\text{max}}^{(ls)}, \quad a_k^{(ls)} = \frac{n}{\sum_{i=1}^{n} \|X_i\|_2^2} \left[ \frac{\lambda_{\text{max}}^{(ls)} - \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right)}{\lambda_{\text{max}}^{(ls)}} \right]^2 - \left( \frac{\|y\|_2^2}{\lambda_{\text{max}}^{(ls)}} \right)^2,
\]

\[
b^{(ls)} = \frac{\|\sum_{i=1}^{n} y_i X_i\|_*}{\lambda_{\text{max}}^{(ls)}}, \quad c^{(ls)} = \frac{1}{\lambda_{\text{max}}^{(ls)}} \left( \sum_{i=1}^{n} y_i - \frac{1}{\lambda_{\text{max}}^{(ls)}} \left( X_i, \sum_{i=1}^{n} y_i X_i \right) \right)^2,
\]

\[
d^{(ls)} = \lambda_{\text{max}}^{(ls)} - \frac{\|y\|_2^2 \cdot \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2}}{\sqrt{n}}, \quad \Delta_k^{(ls)} = \sqrt{(b^{(ls)})^2 + a_k^{(ls)} c^{(ls)}}.
\]

We are ready to give the regularization parameter selection rule for Q-NRM.

**Theorem 4.1** Let \( k \in \{1, 2, \ldots, r\} \) and \( r = \min\{p, q\} \). The sequence of regularization parameters \( \{\lambda_k^{(ls)}\}_{k=1}^{r} \) is defined as

\[
\lambda_k^{(ls)} = \begin{cases} 
\emptyset, & \text{if } (b^{(ls)})^2 + a_k^{(ls)} c^{(ls)} < 0 \\
\max \left\{ 0, \frac{b^{(ls)} + \Delta_k^{(ls)}}{a_k^{(ls)}} \right\}, & \text{if } \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) < d^{(ls)} \\
\frac{c^{(ls)}}{-2b^{(ls)}}, & \text{if } b^{(ls)} < 0, \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) = d^{(ls)} \\
\emptyset, & \text{otherwise,} \\
\left( \lambda_k^{(1)}, \lambda_k^{(u)} \right) \cap [0, +\infty), & \text{if } \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) > d^{(ls)}
\end{cases}
\]

where \( \lambda_k^{(1)} = \frac{b^{(ls)} + \Delta_k^{(ls)}}{a_k^{(ls)}} \), and \( \lambda_k^{(u)} = \frac{b^{(ls)} - \Delta_k^{(ls)}}{a_k^{(ls)}} \).

If \( \lambda \in \bigcup_{i=0}^{k} \lambda_i^{(ls)} \), the solution of Q-NRM satisfies \( \text{rank} \left( B^{(ls)}(\lambda) \right) \leq k - 1 \).

**Proof** From Theorem 3.2, if \( \lambda \) satisfies

\[
\lambda > \frac{\lambda}{\lambda_{\text{max}}^{(ls)}} - \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) + \frac{1}{\lambda_{\text{max}}^{(ls)}} \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2} \cdot \sqrt{\frac{2\text{Gap}^{(ls)}(\lambda)}{n}},
\]

then \( \text{rank} \left( B^{(ls)}(\lambda) \right) \leq k - 1 \). The above inequality of \( \lambda \) is equivalent to

\[
\left( \frac{\lambda_{\text{max}}^{(ls)} - \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right)}{\lambda_{\text{max}}^{(ls)}} \right) \lambda > \frac{1}{\lambda_{\text{max}}^{(ls)}} \sqrt{\sum_{i=1}^{n} \|X_i\|_2^2} \cdot \sqrt{\frac{2\text{Gap}^{(ls)}(\lambda)}{n}}.
\]
To obtain the closed-form of $\lambda$, we square both sides of the above inequality and get that

$$
\left( \lambda_{\text{max}}^{(ls)} - \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) \right) \frac{2}{\lambda_{\text{max}}^{(ls)}} > \sum_{i=1}^{n} \| X_i \|_2^2 \cdot \frac{2 \text{Gap}^{(ls)}(\lambda)}{n}.
$$

Replacing the detailed result of $\text{Gap}^{(ls)}(\lambda)$ and rearranging all terms, we have

$$
\left[ \frac{\lambda_{\text{max}}^{(ls)} - \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right)}{\lambda_{\text{max}}^{(ls)}} \right]^2 \lambda^2 - \sum_{i=1}^{n} \| X_i \|_2^2 \frac{2 \left( \frac{\| y \|_2}{\lambda_{\text{max}}^{(ls)}} \right)^2}{n} \lambda^2 \frac{\lambda_{\text{max}}^{(ls)} - \sigma_k \left( \sum_{i=1}^{n} y_i X_i \right)}{\lambda_{\text{max}}^{(ls)}} \frac{2}{\lambda_{\text{max}}^{(ls)}} > 0.
$$

Dividing both sides of the above inequality by $\frac{\sum_{i=1}^{n} \| X_i \|_2^2}{n}$, we have

$$
a_k^{(ls)} \lambda^2 - 2b^{(ls)} \lambda - c^{(ls)} > 0,
$$

which is a quadratic inequality of $\lambda$. In order to solve it, we need to consider the values of $d_k^{(ls)}$ and $\frac{2b^{(ls)} \pm \sqrt{4d_k^{(ls)}}}{2a_k^{(ls)}}$. We firstly analyze the value of $\frac{2b^{(ls)} \pm \sqrt{4d_k^{(ls)}}}{2a_k^{(ls)}}$. Due to the fact that we consider $\lambda \in \mathbb{R}$, $\lambda_k^{(ls)} \in \emptyset$ if $(b_k^{(ls)})^2 + a_k^{(ls)} c_k^{(ls)} < 0$. Next, we consider that $(b_k^{(ls)})^2 + a_k^{(ls)} c_k^{(ls)} \geq 0$. In this case, the range of $\lambda$ is determined by the value of $a_k^{(ls)}$. Based on the different values of $d_k^{(ls)}$, there are three cases as follows.
(i) For $a_k^{(ls)} > 0$, which equals to $\sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) < d^{(ls)}$, $\lambda$ needs to satisfy that

$$
\lambda > \frac{b^{(ls)} + \Delta^{(ls)}}{a_k^{(ls)}}.
$$

Combining the fact that $\lambda > 0$, so

$$
\lambda > \max \left\{ 0, \frac{b^{(ls)} + \Delta^{(ls)}}{a_k^{(ls)}} \right\}.
$$

(ii) For $a_k^{(ls)} = 0$, which equals to $\sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) = d^{(ls)}$, $\lambda$ needs to satisfy that

$$
-2b^{(ls)} \lambda - c^{(ls)} > 0,
$$

which is a linear inequality of $\lambda$. If $b^{(ls)} > 0$, $\lambda$ needs to satisfy that $\lambda < -\frac{c^{(ls)}}{-2b^{(ls)}} \leq 0$, which is the result of $c^{(ls)} \geq 0$. This result contradicts with the requirement $\lambda > 0$, so $\lambda = \emptyset$ in this case. If $b^{(ls)} = 0$, the inequality $-c^{(ls)} > 0$ is wrong and $\lambda = \emptyset$. If $b^{(ls)} < 0$, $\lambda$ needs to satisfy that $\lambda > -\frac{c^{(ls)}}{-2b^{(ls)}} = \lambda_0^{(ls)}$.

(iii) For $a_k^{(ls)} < 0$, which equals to $\sigma_k \left( \sum_{i=1}^{n} y_i X_i \right) > d^{(ls)}$, $\lambda$ needs to satisfy that $\lambda > 0$ and

$$
\frac{b^{(ls)} + \Delta^{(ls)}}{a_k^{(ls)}} < \lambda < \frac{b^{(ls)} - \Delta^{(ls)}}{a_k^{(ls)}}.
$$

Combining the above arguments, we define $\lambda_k^{(ls)}$ as (9). According to Theorem 3.2, we know that $\text{rank}(B^{*(ls)}(\lambda)) \leq k - 1$ if $\lambda \in \lambda_k^{(ls)}$, and $\text{rank}(B^{*(ls)}(\lambda)) \leq k - 2$ if $\lambda \in \lambda_{k-1}^{(ls)}$. Therefore, if $\lambda \in \bigcup_{i=0}^{k} \lambda_i^{(ls)}$, the solution of Q-NRM satisfies $\text{rank}(B^{*(ls)}(\lambda)) \leq k - 1$.

From Theorem 4.1, we can get a sequence of closed-form regularization parameters and the maximal rank of the solution of Q-NRM under these parameters. Based on the definition of $\lambda_0^{(ls)}$, we know that the set $\bigcup_{i=0}^{k} \lambda_i^{(ls)}$ is not empty, which means that our theorem can always provide a set of $\lambda$ such that $\text{rank}(B^{*(ls)}(\lambda)) \leq k - 1$. If $\lambda_k^{(ls)} = \emptyset$ and $\lambda_{k-1}^{(ls)} \neq \emptyset$, the set $\bigcup_{i=0}^{k} \lambda_i^{(ls)}$ reduces to $\bigcup_{i=0}^{k-1} \lambda_i^{(ls)}$. Therefore, this theorem claims that the solution of Q-NRM satisfies $\text{rank}(B^{*(ls)}(\lambda)) \leq k - 2$ when $\lambda \in \bigcup_{i=0}^{k-1} \lambda_i^{(ls)}$. In the extreme case, $\lambda_i^{(ls)} = \emptyset$ for all $i \in \{1, 2, \ldots, k\}$, the set $\bigcup_{i=0}^{k} \lambda_i^{(ls)}$ reduces to $\{\lambda_0^{(ls)}\}$. Because that 0 is a solution of Q-NRM when $\lambda = \lambda_0^{(ls)}$, $\text{rank}(B^{*(ls)}(\lambda_0^{(ls)})) \leq k - 1$ holds for any $k \in \{1, 2, \ldots, r\}$.

As in Sect. 2, Q-NRM is an efficient tool when the noise has zero mean and constant variance. However, we do not know the distribution of the noise. To deal with this case, H-NRM is a better choice. In the next section, we focus on the regularization parameter selection rule for H-NRM.
4.2 Regularization Parameter Selection Rule for H-NRM

It is well known that the Huber loss function is very important in statistics and machine learning. See, e.g., Huber [13], Sun et al. [25], Elsener and Geer [6]. It is defined as

$$h_\kappa(t) = \begin{cases} \frac{1}{2}t^2, & |t| \leq \kappa \\ \kappa|t| - \frac{1}{2}\kappa^2, & |t| > \kappa. \end{cases}$$

Clearly, the Huber function is differentiable and its gradient is 1-Lipschitz continuous. The conjugate function of the Huber function is

$$h_\kappa^*(\xi) = \max_{t \in \mathbb{R}} \{ t\xi - h_\kappa(t) \} = \begin{cases} \frac{1}{2}\xi^2, & 0 \leq |\xi| \leq \kappa \\ +\infty, & |\xi| > \kappa. \end{cases}$$

H-NRM is given as

$$\min_{B \in \mathbb{R}^{p \times q}} \left\{ \sum_{i=1}^{n} h_\kappa(y_i - \langle X_i, B \rangle) + \lambda \| B \|_* \right\}. \tag{10}$$

The solution of H-NRM is denoted as $B^{*\{h\}}(\lambda)$. Then, its dual form is

$$\max_{\theta \in \mathbb{R}^n} \left\{ \langle y, \theta \rangle - \frac{1}{2} \| \theta \|_2^2 \right\}$$

s.t. $\left\| \sum_{i=1}^{n} \theta_i X_i \right\|_2 \leq \lambda,$

$$-\kappa \mathbf{1} \leq \theta \leq \kappa \mathbf{1}. \tag{11}$$

With the similar way in Sect. 4.1 for Q-NRM, we will give feasible points of (10) and (11), and present the sequence of regularization parameters of H-NRM. The solution of (11) is

$$\theta^{\text{max}\{h\}} = \arg \max_{\theta \in \mathbb{R}^n} \left\{ \langle \theta, y \rangle - \frac{1}{2} \| \theta \|_2^2 \right\} - \kappa \leq \theta_i \leq \kappa, i = 1, 2, \ldots, n,$$

when $B^{*\{h\}}(\lambda) = 0$. The closed-form of $\theta^{\text{max}\{h\}}_i$ is easily obtained as

$$\theta^{\text{max}\{h\}}_i = \begin{cases} y_i, & |y_i| \leq \kappa, \\ \text{sgn}(y_i) \kappa, & |y_i| > \kappa, \quad i = 1, 2, \ldots, n. \end{cases}$$

Let $\tau_i = y_i \delta_{|y_i| \leq \kappa}(y_i) + \text{sgn}(y_i) \cdot \kappa \cdot \delta_{|y_i| > \kappa}(y_i)$ and $\tau = (\tau_1, \ldots, \tau_n)^T$. Then $\theta^{\text{max}\{h\}} = \tau$ and the lower bound of the regularization parameter that enforces the solution of H-NRM being zero is $\lambda^{(h)}_{\text{max}} = \left\| \sum_{i=1}^{n} \theta^{\text{max}\{h\}}_i X_i \right\|_2 = \left\| \sum_{i=1}^{n} \tau_i X_i \right\|_2$. For any $0 < \lambda < \lambda^{(h)}_{\text{max}},$...
the feasible points of (10) and (11) are set as follows.

\[
\theta_0^{(h)}(\lambda) = \frac{\lambda \theta^{\text{max}(h)}}{\lambda^{(h)}_{\text{max}}} = \frac{\lambda \tau}{\lambda^{(h)}_{\text{max}}} , \quad B_0^{(h)}(\lambda) = \frac{\sum_{i=1}^{n} \theta_i^{0(h)}(\lambda) X_i}{\lambda} = \frac{\sum_{i=1}^{n} \tau_i X_i}{\lambda^{(h)}_{\text{max}}} .
\]

Therefore, the dual solution is contained in the set

\[
\left\{ \theta \left| \| \theta - \theta_0^{(h)}(\lambda) \|_2 \leq \sqrt{\frac{2 \text{Gap}^{(h)}(\lambda)}{n}} \right. \right\},
\]

where \( \text{Gap}^{(h)}(\lambda) \) is

\[
\sum_{i=1}^{n} h_k(y_i - \frac{(X_i, \sum_{i=1}^{n} \tau_i X_i)}{\lambda^{(h)}_{\text{max}}}) + \frac{\lambda \tau}{\lambda^{(h)}_{\text{max}}} \| \sum_{i=1}^{n} \tau_i X_i \|_* - (\frac{\lambda}{\lambda^{(h)}_{\text{max}}}) (y, \tau) - \frac{1}{2} (\frac{\lambda}{\lambda^{(h)}_{\text{max}}})^2 \| \tau \|_2^2 .
\]

Moreover, we define some new notations.

\[
\lambda^{(h)}_0 = \lambda^{(h)}_{\text{max}}, \quad d^{(h)}_k = \frac{n}{\sum_{i=1}^{n} \| X_i \|_2^2} \left( \frac{\lambda^{(h)}_{\text{max}} - \sigma_k \left( \frac{n}{\sum_{i=1}^{n} \tau_i X_i} \right) \| \tau \|_2^2}{\lambda^{(h)}_{\text{max}}} \right),
\]

\[
b^{(h)} = \frac{\| \sum_{i=1}^{n} \tau_i X_i \|_*}{\lambda^{(h)}_{\text{max}}} - \| \tau \|_2^2 , \quad c^{(h)} = \sum_{i=1}^{n} \left( \tau_i - \frac{1}{\lambda^{(h)}_{\text{max}}} \left( X_i, \sum_{i=1}^{n} \tau_i X_i \right) \right) ,
\]

\[
d^{(h)} = \lambda^{(h)}_{\text{max}} - \frac{\| \tau \|_2 \cdot \sqrt{\sum_{i=1}^{n} \| X_i \|_2^2}}{\sqrt{n}} , \quad \Delta^{(h)}_k = \sqrt{(b^{(h)})^2 + a^{(h)}_k c^{(h)}} .
\]

The next theorem shows a sequence of regularization parameters and helps to select the suitable regularization parameter when the upper bound of the rank of the solution is known. Because the proof of the next theorem is similar to that of Theorem 4.1, we omit it.

**Theorem 4.2** Let \( k \in \{1, 2, \ldots, r\} \) and \( r = \min\{p, q\} \). The sequence of regularization parameters \( \{\lambda_k^{(h)}\}_{k=1}^{r} \) is defined as

\[
\lambda_k^{(h)} = \left\{ \begin{array}{ll}
\emptyset, & (b^{(h)})^2 + a_k^{(h)} c^{(h)} < 0 \\
\max \left\{ 0, \frac{b^{(h)} + \Delta^{(h)}_k}{a_k^{(h)}} \right\}, & \sigma_k \left( \frac{n}{\sum_{i=1}^{n} \tau_i X_i} \right) < d^{(h)} \\
\left\{ \frac{c^{(h)}}{2b^{(h)}}, \quad b^{(h)} < 0, \right\} & \sigma_k \left( \frac{n}{\sum_{i=1}^{n} \tau_i X_i} \right) = d^{(h)} \\
\emptyset, & \sigma_k \left( \frac{n}{\sum_{i=1}^{n} \tau_i X_i} \right) > d^{(h)} \\
\left( \lambda_k^{(2)}, \lambda_k^{(2)} \right) \cap (0, +\infty), & \sigma_k \left( \frac{n}{\sum_{i=1}^{n} \tau_i X_i} \right) > d^{(h)}
\end{array} \right.
\]
Fig. 1 The signal shapes in this paper. These shapes are all black-and-white pictures and are transformed into matrices with elements 0 and 1, where the white parts are represented by 0 and the black parts are 1. These shapes are all $64 \times 64$ matrices. The rank of these signal shapes are 19, 12, 2 and 28 respectively.

\[
\begin{align*}
\lambda_k^{(2)} &= \frac{b^{(h)} + \Delta_k^{(h)}}{a_k^{(h)}} \quad \text{and} \quad \lambda_k^{(2)} = \frac{b^{(h)} - \Delta_k^{(h)}}{a_k^{(h)}}.
\end{align*}
\]

If $\lambda \in \bigcup_{i=0}^{k} \lambda_i^{(h)}$, the solution of H-NRM satisfies that $\text{rank} \left( B^{s(h)}(\lambda) \right) \leq k - 1$.

5 Numerical Experiments

In this section, we report some numerical results to illustrate the efficiency of the regularization parameter selection rule. All experiments were carried on a PC with Inter(R) Core(TM) i5-8250U CPU @ 1.60 GHz and 8.00GB RAM using MATLAB version R2018b.

We introduce the idea of the traditional cross validation to choose the optimal $\lambda$ for NRM. For the possible interval $[0, \lambda_{\text{max}}]$, where $\lambda_{\text{max}}$ is defined in Lemma 3.1, the cross validation needs to select some values in this interval and solve NRM with these values. By comparing these solutions with a certain criterion, the optimal $\lambda$ can be selected. These criteria can be test error or information criteria, such as AIC, BIC and so on.

Here, we choose some signal shapes from MPEG-7 1, which are showed in Fig. 1. These pictures were analyzed in Zhou and Li [33], where they proposed the Nesterov method to solve (7). In their paper, the cross validation was used to choose the optimal regularization parameter. So, we embed our regularization parameter rule into their code and report the computational effects. For every shape, we randomly simulate matrices $X_i$ ($i \in \{1, 2, \ldots, 500\}$) with elements obeying the uniform distribution on $[0, 0.01]$, and the noise $\epsilon_i \sim N(0, 0.1)$. Then, $y_i$ is obtained by

\[
y_i = \langle X_i, B \rangle + \epsilon_i,
\]

where $B$ is the numerical matrix of the shape. To perform the cross validation, we randomly partition the data set into 10 folds, where 1 fold is the test data and the others are the training data. We denote the training index and test index as $I_{\text{train}}$ and $I_{\text{test}}$, respectively.

---

1 http://www.dabi.temple.edu/shape/MPEG7/index.html.
Table 1 Comparison of the traditional cross validation and our results with int = 500, when the upper bound of the rank of the solution is known. Here, int = 500 means that the number of $\lambda$ in $[0, \lambda_{\text{max}}]$ for the cross validation is 500

| Name     | $\lambda_{(ls)}$ | $\lambda_{\text{cv}}$ | $\lambda_{\text{rcv}}$ | Time(s)  | RMSE  | RMSE$_{\text{cv}}$ | RMSE$_{\text{rcv}}$ |
|----------|------------------|------------------------|-------------------------|----------|-------|---------------------|---------------------|
| device1-1 | 19.8479          | 0.0397                 | 13.8141                 | 23.0994  | 11.6624 | 0.0449              | 0.0995              |
| device6-20 | 23.5412          | 0.0670                 | 4.5581                  | 23.1243  | 16.3582 | 0.0493              | 0.0873              |
| device8-20 | 19.4122          | 0.1941                 | 14.5427                 | 21.2577  | 10.3726 | 0.0459              | 0.1039              |
| lizard-3   | 27.6182          | 0.1105                 | 8.0540                  | 22.8897  | 18.2546 | 0.0544              | 0.0763              |

Suppose the upper bound of the rank of the solution is known as $k - 1$, according to Theorem 3.2, the possible interval can be shrunk into

$$F_k = \left[ \sigma_k \left( \sum_{i=1}^{n} \theta_i(\lambda) X_i \right) + \sqrt{\frac{2\text{Gap}(\lambda)}{\alpha} \cdot \sum_{i=1}^{n} \| X_i \|_2^2}, \lambda_{\text{max}} \right].$$

In order to present the efficiency of our rule, we only choose the regularization parameter values in the $F_k$ for the cross validation and decide the optimal $\lambda$. In the following results, we select 500 values of $\lambda$ that are equally spaced on the $\lambda/\lambda_{(ls)}$ scale from 0.01 to 1. Let the test error be the criterion. $\lambda_{\text{cv}}$ means the optimal regularization parameter in (7) under the traditional cross validation, and $\lambda_{\text{rcv}}$ means the optimal regularization parameter under our regularization parameter selection rule. $t_{\text{cv}}$ means the computational time for the traditional cross validation and $t_{\text{rcv}}$ means the computational time for the cross validation that is embedded our regularization parameter selection rule. Denote $B^{\lambda_{\text{cv}}}$ and $B^{\lambda_{\text{rcv}}}$ as the solution of (7) under $\lambda_{\text{cv}}$ and $\lambda_{\text{rcv}}$, respectively. To measure the performance of $\lambda_{\text{cv}}$ and $\lambda_{\text{rcv}}$, we use RMSE$_{\text{cv}}$ and RMSE$_{\text{rcv}}$, where

$$\text{RMSE}_{\text{cv}} = \frac{\| Y_{\text{test}} - \hat{y}_{\text{cv}} \|_2}{\sqrt{|I_{\text{test}}|}} \quad \text{with} \quad \hat{y}_{\text{cv}} = \langle X_i, B^{\lambda_{\text{cv}}} \rangle, \quad i \in I_{\text{test}}$$

and

$$\text{RMSE}_{\text{rcv}} = \frac{\| Y_{\text{test}} - \hat{y}_{\text{rcv}} \|_2}{\sqrt{|I_{\text{test}}|}} \quad \text{with} \quad \hat{y}_{\text{rcv}} = \langle X_i, B^{\lambda_{\text{rcv}}} \rangle, \quad i \in I_{\text{test}}.$$

From Table 1, we can get the following conclusions. (i) Under the help of our regularization parameter selection rule, the computational time for the selection of the regularization parameter is reduced. For device1-1 and device8-20, $t_{\text{rcv}}$ is smaller than half of $t_{\text{cv}}$. (ii) Although the optimal $\lambda$ of our screening rule is larger than that of the traditional cross validation, which is because that our rule screens out smaller values of $\lambda$ to obtain the low rank solution, the performance of the optimal $\lambda$ under our rule is almost same with that of the cross validation.
**Table 2** Comparison of the traditional cross validation and our results with int = 500, when the upper bound of the rank of the solution is unknown. Here, int = 500 means that the number of $\lambda$ in $[0, \lambda_{\text{max}}]$ for the cross validation is 500.

| int = 500 | Time(s) | Test error | AIC | BIC |
|-----------|---------|------------|-----|-----|
| device1-1 | CV 16.4563 | 0.3578 | 0.0133 | 0.0357 | 265.0799 | 1.4311 | 807.4160 |
|           | RS 0.5594 | 6.2611 | 0.0242 | 6.2611 | 266.6745 | 6.2611 | 808.2605 |
| device6-20 | CV 16.1499 | 0.4253 | 0.0154 | 0.4253 | 265.2126 | 0.4253 | 807.5876 |
|           | RS 0.5767 | 7.8685 | 0.0307 | 7.8685 | 267.8174 | 7.8685 | 809.3540 |
| device8-20 | CV 15.9713 | 0.3493 | 0.0171 | 0.3493 | 264.9881 | 1.3973 | 807.3165 |
|           | RS 0.5739 | 6.1134 | 0.0288 | 6.1134 | 266.4905 | 6.1134 | 808.0765 |
| lizard-3  | CV 17.4973 | 0.4975 | 0.0183 | 0.4975 | 265.3482 | 0.4975 | 807.7232 |
|           | RS 0.9800 | 9.4521 | 0.0384 | 9.4521 | 269.1887 | 9.4521 | 810.7006 |

Suppose that the upper bound of the rank of the solution is unknown. To evaluate the regularization parameter selection rule in this paper, we calculate the sequence of the regularization parameters and choose one of them as the optimal parameter under a criterion. This idea is to select some parameters that are guaranteed to lead different upper bounds of the rank of the solution. To verify this idea carefully, we compare it with the cross validation under three criteria, that is test error, AIC and BIC. One can see definitions of AIC and BIC in Zhou and Li [33]. The smaller of these criteria, the better of the regularization parameter. The cross validation and our rule are performed on the training data and evaluated on the test data. In Table 2, CV presents the cross validation and RS presents our regularization parameter selection rule Theorem 4.1. $\lambda_{\text{opt}}$ means the optimal regularization parameter. AIC$_{\text{opt}}$ and BIC$_{\text{opt}}$ represent the smallest value of AIC and BIC, respectively.

From Table 2, we can conclude the following results. (i) Our rule can reduce the computational time for the selection of the regularization parameter. Comparing to the traditional cross validation, our rule can reduce the computational time by orders of magnitude. (ii) The optimal $\lambda$ under our rule is larger than that of cross validation. This is the result that our rule is safe and the sequence of the regularization parameters is larger than the true one. (iii) Under different criteria, the optimal $\lambda$ of our rule is stable. For device1-1 and device8-20, under test error, AIC and BIC, $\lambda_{\text{opt}}$ for the cross validation varies, while $\lambda_{\text{opt}}$ for our rule is same.

### 6 Conclusion

With the help of the duality theory, we propose a novel regularization parameter selection rule for NRM. If the rank of the solution has an upper bound, this rule provides a safe set of the regularization parameter. Moreover, we apply this method to the nuclear norm regularized quadratic minimization and the nuclear norm regularized Huber minimization. We get the sequence of closed-form regularization parameters of these two problems. Finally, we illustrate the efficiency of the regularization parameter.
selection rule on some signal shapes, which indicate that our rule can reduce the computational time for the selection of the regularization parameter.

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

References

1. Bottou, L., Curtis, E.F., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. 60(2), 223–311 (2018)
2. Cai, T.T., Zhang, A.: Sharp RIP bound for sparse signal and low-rank matrix recovery. Appl. Comput. Harmon. Anal. 35(1), 74–93 (2013)
3. Chen, S., Donoho, D., Saunders, M.: Atomic decomposition for basis pursuit. SIAM J. Sci. Comput. 20(1), 33–61 (1998)
4. Chen, C., Wei, C., Wang, F.Y.: Low-rank matrix recovery with structural incoherence for robust face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2618–2625 (2012)
5. Davenport, A.M., Romberg, J.: An overview of low-rank matrix recovery from incomplete observations. IEEE J-STSP 10(4), 608–622 (2016)
6. Elsener, A., Geer, S.: Robust low-rank matrix estimation. Ann. Stat. 48, 3481–3509 (2018)
7. Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space (with discussion). J. R. Stat. Soc. B 70, 849–911 (2008)
8. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis, Stanford University (2002)
9. Fazel, M., Pong, T.K., Sun, D., Tseng, P.: Hankel matrix rank minimization with applications to system identification and realization. SIAM J. Matrix Anal. Appl. 34(3), 946–977 (2013)
10. Ghaoui, E.L., Viallon, V., Rabbani, T.: Safe feature elimination in sparse supervised learning. Pac. J. Optim. 8(4), 667–698 (2012)
11. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms. Springer, Berlin (1993)
12. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2013)
13. Huber, P.J.: Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. 1, 799–821 (1973)
14. Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear norm penalization and optimal rates for noisy low rank matrix completion. Ann. Stat. 39, 2302–2329 (2011)
15. Kong, L.C., Tunçel, L. and Xiu, N.H.: s-goodness for low-rank matrix recovery. Abstr. Appl. Anal. 101974, 9 pages (2013)
16. Kuang, Z.B., Geng, S.N., Page, D.: A screening rule for l_1-regularized ising model estimation. Adv. Neural. Inf. Process. Syst. 30, 720–731 (2017)
17. Lee, S., Gornitz, N., Xing, E.P., Heckerman, D., Liptert, C.: Ensembles of Lasso screening rules. IEEE Trans. Pattern Anal. 40(12), 2841–2852 (2018)
18. Liu, Z., Vandenberghc, L.: Interior-point method for nuclear norm approximation with application to system identification. SIAM J. Matrix Anal. Appl. 31(3), 1235–1256 (2009)
19. Lu, Z., Monteiro, R.D.C., Yuan, M.: Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Math. Program. 131, 163–194 (2010)
20. Ndiaye, E., Fercqo, O., Gramfort, A., Salmon, J.: Gap safe screening rules for sparsity enforcing penalties. J. Mach. Learn. Res. 18, 1–33 (2017)
21. Negahban, S., Wainwright, M.J.: Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Stat. 39(2), 1069–1097 (2011)
22. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. SIAM Rev. 52(3), 471–501 (2010)
23. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
24. Rohde, A., Tsybakov, A.B.: Estimation of high-dimensional low-rank matrices. Ann. Stat. 39, 887–930 (2011)
25. Sun, Q., Zhou, W., Fan, J.: Adaptive Huber regression. J. Am. Stat. Assoc. 115(529), 254–265 (2020)
26. Tibshirani, R., Bien, J., Hastie, T., Simon, N., Taylor, J., Tibshirani, R.J.: Strong rules for discarding predictors in lasso-type problems. J. R. Stat. Soc. B 74(2), 1–22 (2012)
27. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58, 267–288 (1996)
28. Wang, J., Wonka, P., Ye, J.: Lasso screening rules via dual polytope projection. J. Mach. Learn. Res. 16, 1063–1101 (2015)
29. Watson, G.A.: Characterization of the subdifferential of some matrix norms. Linear Algebra Appl. 170, 33–45 (1992)
30. Xiang, Z.J., Wang, Y., Ramadge, J.P.: Screening tests for Lasso problems. IEEE. Trans. Pattern Anal. 5(39), 1008–1027 (2017)
31. Yuan, M., Ekici, A., Lu, Z., Monteiro, R.D.C.: Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. Ser. B 69, 329–346 (2007)
32. Zhao, B., Justin, P.H., Cornelius, B., Liang, Z.P.: Low rank matrix recovery for real-time cardiac MRI. In: International Symposium on Biomedical Imaging, pp. 996–999 (2010)
33. Zhou, H., Li, L.X.: Regularized matrix regression. J. R. Stat. Soc. B 76(2), 463–483 (2014)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.