Machine Learning (2023) 112:2723–2760
https://doi.org/10.1007/s10994-023-06325-w

Robust matrix estimations meet Frank–Wolfe algorithm

Naimin Jing[1,3] · Ethan X. Fang[2] · Cheng Yong Tang[1]

Received: 7 May 2021 / Revised: 13 April 2022 / Accepted: 26 February 2023 /
Published online: 5 April 2023

Abstract

We consider estimating matrix-valued model parameters with a dedicated focus on their robustness. Our setting concerns large-scale structured data so that a regularization on the matrix's rank becomes indispensable. Though robust loss functions are expected to be effective, their practical implementations are known difficult due to the non-smooth criterion functions encountered in the optimizations. To meet the challenges, we develop a highly efficient computing scheme taking advantage of the projection-free Frank–Wolfe algorithms that require only the first-order derivative of the criterion function. Our methodological framework is broad, extensively accommodating robust loss functions in conjunction with penalty functions in the context of matrix estimation problems. We establish the non-asymptotic error bounds of the matrix estimations with the Huber loss and nuclear norm penalty in two concrete cases: matrix completion with partial and noisy observations and reduced-rank regressions. Our theory demonstrates the merits from using robust loss functions, so that matrix-valued estimators with good properties are achieved even when heavy-tailed distributions are involved. We illustrate the promising performance of our methods with extensive numerical examples and data analysis.

Keywords Frank–Wolfe algorithms · Huber loss · Matrix-valued parameters · Robust statistical methods · Non-asymptotic properties · Non-smooth criterion function

Editor: Pradeep Ravikumar.

✉ Cheng Yong Tang
yongtang@temple.edu

Naimin Jing
naimin.jing@merck.com

Ethan X. Fang
xingyuan.fang@duke.edu

[1] Department of Statistics, Operations, and Data Science, Fox School of Business, Temple University, Philadelphia, PA, USA

[2] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

[3] Present Address: Biostatistics and Research Decision Sciences, Merck & Co., Inc, Kenilworth, NJ, USA

1 Introduction

Massive data with informative structures from the data collection processes are becoming increasingly available in many data-enabled areas. Examples include those from FMRI, electroencephalogram (EEG), and tick-by-tick financial trading records of many assets. Methodologically for multivariate data analysis, matrices as the model parameters are commonly analyzed in the core step(s) of many popular approaches including the principal component analysis, canonical correlation analysis (Anderson, 2003), Gaussian graphical model analysis (Lauritzen, 1996), reduced-rank regression (Reinsel & Velu, 1998), sufficient dimension reduction (Cook, 2009), and many others.

Structural information—our foremost consideration in this study—is indispensable in solving many matrix estimation problems with large-scale data. For matrix-valued model parameters, a class of methods imposes restrictions on the rank of the targeted matrix. In matrix completion with partial and noisy observations, for example, without such structural information, successfully recovering the signal is not possible. For multi-response regression problems, structural information is vital for both methodological development and practical implementation for drawing informative conclusions. Constraining the rank of the parameter matrix in multi-response regression leads to the conventional reduced rank regression (Reinsel & Velu, 1998).

Our primary goal in this study is to investigate robustness when estimating matrices with large-scale data and structural information. Robustness is a foundational concern in current data-enabled investigations. During massive data collection processes, observations of heterogeneous quality are inevitable, and even erroneous records are common. On one hand, due to the huge size of the data in modern large-scale investigations, validations and error corrections become too daunting to be practical. Robust statistical methods in these scenarios are thus highly desirable. On the other hand, however, in many existing methods, though being convenient, commonly applied criterion functions including the squared loss and the negative log-likelihood are unfortunately not robust to the violations of the model assumptions in the aforementioned practical reality.

We are thus motivated to consider robustness in the context with structural information, which is incorporated by constraining the rank of the matrix-valued model parameters. The foremost challenge in this scenario is the fundamental computational difficulty. One source contributing to the difficulty roots in the fact that constraining a matrix's rank results in a non-convex problem. As a rare example in reduced-rank multivariate regression, an analytic solution is available despite the non-convexity; see (Reinsel & Velu, 1998). Unfortunately when considerations are broader, such a convenience generally no longer exists; and how to solve optimization problems with rank constraints is generally difficult. To meet the challenge, a convex relaxation of the problem leads to regularizing the nuclear norm of the matrix-valued model parameter. From the statistical perspective, numerous works (Candès & Tao, 2010; Negahban & Wainwright, 2011; Agarwal et al., 2012) have studied the theoretical properties of this type of estimators constructed with the nuclear norm relaxation, and have proved that the resulting estimator achieves optimal or near-optimal statistical properties under different settings. Additional to the non-convexity, consideration of robustness is further contributing to the computational difficulty. Resorting to robust loss functions is a traditional class of influential methods for establishing more robust statistical methods; see Huber (2004) and Hampel et al. (2011). Though demonstrated effective in conventional statistical analysis, substantial difficulties arise when handling large-scale modern complex data-enabled problems. Computationally, in particular, their applications

encounter major challenges because robust loss functions are not smooth whose second-order derivatives do not exist. Analytically, establishing the statistical properties of the matrix estimations is challenging in this scenario too, because the impacts from possibly heavy-tailed errors are involved in studying large-scale problems. Existing methods using the squared loss or the negative log-likelihood as the loss functions require the noises to be sub-Gaussian in order to handle high-dimensional data. Robust methods can accommodate noises with heavier tails than sub-Gaussian; meanwhile, the capacity for handling high-dimensional data remains desirable.

There has been an active recent development in robust statistical methods with high-dimensional data; see, for example, Loh (2017), Zhou et al. (2018), Sun et al. (2020), and reference therein. Recently, there has been increasing interest in investigating robust methods for matrix-valued model parameters. She and Chen (2017) studied the robust reduced-rank regression in a scenario concerning outliers. They define the estimator as the minimizer of a non-convex optimization problem, establish theoretical error bounds, and propose to apply an iterative algorithm that alternatively solves for two parts of the model parameters in their setting. Due to the nonconvexity, their algorithm does not guarantee the convergence to the minimum. Wong and Lee (2017) studied matrix completion with Huber loss. Their algorithm is developed by iteratively projecting non-robust matrix estimators, which is computationally demanding with many projection operations required. Elsener and van de Geer (2018) investigated robust matrix completion with the Huber loss function and nuclear norm penalization. The computation algorithms in Elsener and van de Geer (2018) involved a soft-thresholding step for singular values. This works well when the solution is of exact low rank. However, when the solution is of approximately low rank, or of modestly higher rank, such a step becomes computationally demanding. As pointed out in She and Chen (2017), efficient algorithms are desirable for solving optimization problems with rank constraints and robust loss functions.

We attempt our study with a foremost consideration on an efficient computing scheme for solving large-scale statistical problems with robustness. In particular, we aim to develop efficient first-order algorithms by building a scheme with Frank–Wolfe-type algorithms for robust matrix estimation problems. The Frank–Wolfe algorithm is a first-order method and is drawing considerable attention recently (Jaggi, 2013; Lacoste-Julien & Jaggi, 2015; Freund & Grigas, 2016; Freund et al., 2017; Kerdreux et al., 2018; Swoboda & Kolmogorov, 2019). The key advantage of the Frank–Wolfe algorithms is their freedom from the required projections in most proximal-type algorithms. In addition, as we shall see in our algorithms in Sect. 2, for matrix estimation problems, in each iteration, the Frank–Wolfe algorithm only requires computing the top one leading singular vectors, which can be conducted efficiently even for huge-size problems. These merits make Frank–Wolfe-type algorithms particularly appealing for solving large-scale robust low-rank matrix estimation problems.

Our study makes two main contributions. Foremost, we develop a new computation scheme for robust matrix estimation and demonstrate that the first-order optimization technique makes solving large-scale robust estimation problems practically convenient. We show extensively that our framework is broadly applicable, covering general robust loss functions including those used in median and quantile regression; see Sect. 2. Second, our theoretical analysis reveals the benefit from using robust loss functions and rank constraints. Our non-asymptotic results demonstrate that our framework can accommodate high-dimensional data. For matrix completion and reduced-rank regression, the resulting matrix-valued estimator works satisfactorily even when the model error distributions are heavy-tailed.

The rest of this article is organized as follows. Section 2 elaborates a concrete framework using the Frank–Wolfe algorithm to solve robust matrix estimation problems. We present matrix completion and reduced-rank regression with various robust loss functions. Section 3 justifies the validity of our method with theory on the algorithm convergence and error bounds of the resulting estimators. Section 4 presents extensive numerical examples demonstrating the promising performance of our methods.

For a generic matrix A, we denote by $A^\top$ its transpose, $\sigma_1(A)$ its largest singular value, $\|A\|_*$ its nuclear norm, and $\|A\|_F$ its Frobenius norm. Let $\langle A, B\rangle = \mathrm{trace}(A^\top B)$ for $A, B \in \mathbb{R}^{p\times q}$. We denote by $\Theta \in \mathbb{R}^{p\times q}$ a generic matrix-valued model parameter. In this study, we focus on two concrete cases. In one case, $\Theta = M$ where M is the signal to be recovered in the matrix completion problem with a single copy of partial and noisy observations; the other one is $\Theta = C$ where C is the matrix-valued coefficients in the multi-response regression problem. Furthermore, we show that our framework broadly applies in solving a general class of problems.

2 Methodology

2.1 Matrix completion

We consider the matrix completion problem first. In this setting, one observes a noisy subset of all entries of a matrix $M \in \mathbb{R}^{p\times q}$, which is the model parameter of interest. Let the set of observed entries be $\Omega = \{(i_t, j_t)\}_{t=1}^n$, where $i_t \in \{1, \ldots, p\}$, $j_t \in \{1, \ldots q\}$, and denote by $X_{i_t j_t}, (i_t, j_t) \in \Omega$, the corresponding noisy observations such that

$$X_{i_t j_t} = M_{i_t j_t} + \xi_t, \quad t = 1, \ldots, n.$$

We assume that ξ_t's are independent and identically distributed random variables with mean zero.

To effectively recover M with a single copy of partial and noisy observations over Ω, one popular approach is to assume that the underlying true matrix, denoted by M^*, is of low-rank that $\mathrm{rank}(M^*) \le r$ for some $r \le \min(p, q)$. Then one can estimate M^* by solving a constrained optimization problem by minimizing the objective function $(2n)^{-1}\sum_{t=1}^n \ell(X_{i_t j_t} - M_{i_t j_t})$ over M, subject to $\mathrm{rank}(M) \le r$ for some loss function $\ell(\cdot)$. Since the rank constraint is non-convex, solving the optimization is generally not tractable. To obtain a practical solution, a common strategy is relaxing the rank constraint to the convex nuclear norm constraint.

The Huber loss function leads to robust estimators because its design alleviates the excessive contribution from a data point that is extremely deviated from the fit. Practically, the Huber loss performs promisingly when handling a substantial portion of noisy observations whose distribution can be heavy-tailed; see Huber (2004).

By applying the Huber loss with a constraint on nuclear norm, we consider the following robust matrix completion problem:

$$\min_{M\in\mathbb{R}^{p\times q}} L_\eta(M) := \frac{1}{2n}\sum_{t=1}^n \ell_\eta(X_{i_t j_t} - M_{i_t j_t}), \ \text{subject to} \|M\|_* \le \lambda \tag{1}$$

where $\ell_\eta(\cdot)$ is the classical Huber loss function:

$$\ell_\eta(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le \eta \\ \eta \cdot (|x| - \frac{1}{2}\eta) & \text{otherwise.} \end{cases} \tag{2}$$

Here η is the tuning parameter of the Huber loss, and λ is the tuning parameter regularizing the nuclear norm of M. In our numerical studies, we choose the tuning parameters by applying the cross-validation.

Since ℓ_η is not smooth, those methods commonly applied in solving ℓ_2-loss problems—requiring second-order derivatives—do not directly apply. Computing optimization problem (1) is generally hard; see the discussion in She and Chen (2017). Efficient algorithms for solving (5) are lacking; the primary difficulty is due to the absence of the second-order derivative of the Huber loss. It is even more challenging to minimize the Huber loss on a restricted low-rank region, and to achieve the computational efficiency with large-scale data. More broadly, non-smooth criterion functions are commonly the case with general robust loss functions, with prominent examples including the least absolute deviation loss of the median regression, check loss of the quantile regression, and Tukey's biweight loss besides the aforementioned Huber loss.

To address the computational difficulty when handling large-scale problems with robust loss functions, we propose to apply the Frank-Wolfe algorithm to solve this problem. The Frank–Wolfe algorithm has been particularly powerful for convex optimizations. As a first-order approach that requires no second-order derivative of the criterion function, it is particularly powerful for solving problems with non-smooth loss functions, which is exactly the case for our problem (1). Briefly speaking, the Frank–Wolfe algorithm pursues some constrained approximation of the gradient—the first-order derivative of the criterion function evaluated at a given value. The algorithm runs iteratively, with the optimization proceeding along the direction as identified by the approximation of the gradient. Therefore, the Frank–Wolfe algorithm is practically appealing, as one has the opportunity to best exploit some constrained approximation that can be computed efficiently. For a detailed account of the Frank–Wolfe algorithms and recent advances in the area, we refer to Freund and Grigas (2016), Freund et al. (2017), and references therein.

Concretely in our setting, we develop an algorithm that runs iteratively. Specifically, at the $(k+1)$-th iteration with $M^{(k)}$ from the previous step, the matrix-valued gradient of (1): $\nabla \mathrm{L}(M^{(k)}) \in \mathbb{R}^{p\times q}$ is analytically calculated by

$$\begin{aligned} \nabla \mathrm{L}(M^{(k)}) = \frac{1}{2n}\sum_{t=1}^{n} J_t \big[& (M^{(k)}_{i_t j_t} - X_{i_t j_t}) 1(|M^{(k)}_{i_t j_t} - X_{i_t j_t}| \le \eta) \\ & + \eta \operatorname{sign}(M^{(k)}_{i_t j_t} - X_{i_t j_t}) 1(|M^{(k)}_{i_t j_t} - X_{i_t j_t}| > \eta) \big], \end{aligned} \tag{3}$$

where J_t is a matrix with $J_{t,i_t j_t} = 1$ and all the other entries 0, $1(\cdot)$ is the indicator function, and $\operatorname{sign}(x) = 1$ if x is positive and -1 otherwise. Hence, evaluating the gradient can be done efficiently, and it is a scalable process that can be efficiently distributed if multiple computing units are available. Then, the Frank–Wolfe algorithm suggests computing a descent direction in the $(k+1)$-th iteration:

$$V^{(k+1)} \leftarrow \operatorname*{argmin}_{V} \langle \nabla \mathrm{L}(M^{(k)}), V \rangle, \text{ subject to } \|V\|_* \le \lambda$$

In this step, a key observation is that

$$V^{(k+1)} = -\lambda\, u_1 v_1^\top, \tag{4}$$

where u_1 and v_1 are the leading left and right singular vectors of $\nabla \mathcal{L}_\eta(M^{(k)})$. The required singular decomposition can be computed efficiently by an existing algorithm that is implemented in the standard "PROLACK" package in Matlab. Then, we conduct a descent step to update $M^{(k)}$ by

$$M^{(k+1)} \leftarrow M^{(k)} + \alpha_{k+1}\left(V^{(k+1)} - M^{(k)}\right),$$

where $\alpha_{k+1} \in [0, 1]$ is a pre-specified step-size. For example, $\alpha_{k+1} = 1/(k+3)$ guarantees convergence to an optimal solution. Meanwhile, line search is viable, and there are various ways to further accelerate this algorithm.

Intuitively, the updating direction in Equation (4) is viewed as the best rank-one approximation of the gradient matrix (3). Further, if we view the vector v_1 as the direction corresponding to the first principal component of the columns of M, then formula (4) is essentially a column-wise update along this direction, with the step sizes proportional to the components in the vector. From this perspective, the update formula (4) can also be viewed as a computationally efficient matrix-valued coordinate descent along the direction u_1. Since the objective function (1) is convex, such an update progressing along the gradient direction ensures that the criterion function converges, approaching the minimum.

We summarize the algorithm in Algorithm 1.

Algorithm 1 Frank-Wolfe Algorithm for Robust Matrix Completion

Input: $\{X_{i_t,j_t}\}_{t=1}^n$, η, $\{\alpha_k\}_{k\in\mathbb{N}}$, λ, $M^{(0)}$, $k = 0$
Output: $\hat{M}$.

```
while stopping criterion is not met do
    u_1 ← left leading singular vector of ∇L_η(M^(k))
    v_1 ← right leading singular vector of ∇L_η(M^(k))
    V^(k+1) ← −λ · u_1 v_1^⊤
    M^(k+1) ← M^(k) + α_{k+1}(V^(k+1) − M^(k))
    k ← k + 1
end while
M̂ ← M^(k)
```

2.2 Reduced-rank regression

In our second concrete problem with matrix-valued model parameters, we consider a multivariate linear regression

$$y_{ij} = x_i^\top c_j + \xi_{ij}, \qquad \text{for } i = 1, \ldots, n,\ j = 1, \ldots, q,$$

where ξ_{ij}'s are model errors. We assume that ξ_{ij}'s are independent and identically distributed random variables with mean zero. Then, we have in a matrix form

$$Y = XC + \Xi,$$

where $Y = [y_{ij}]_{n\times q}$, $X = [x_{ij}]_{n\times p} = [x_1, \ldots, x_n]^\top$, $C = [c_1, \ldots, c_q] \in \mathbb{R}^{p\times q}$, and $\Xi = [\xi_{ij}]_{n\times q}$.

In this setting, one may opt to restrict the rank of C, $\mathrm{rank}(C) \le r\ (r \le \min(p, q))$—leading to the conventional reduced-rank regression (Reinsel & Velu, 1998). Also by relaxing the rank constraint with the nuclear norm, we consider the estimation problem as

$$\min_{C\in\mathbb{R}^{p\times q}} L_\eta(C) := \sum_{i=1}^{n}\sum_{j=1}^{q} \ell_\eta(y_{ij} - x_i^\top c_j), \text{ subject to } \|C\|_* \le \lambda, \tag{5}$$

where c_j denotes the j-th column of C, and $\ell_\eta(\cdot)$ is the Huber loss function with parameter η.

Again, to address the computational challenges, analogous to problem (1), we propose to solve problem (5) also by applying Frank–Wolfe algorithm iteratively with the steps described as follows. Denote by $C^{(k)}$ the solution after the k-th iteration. At the $(k+1)$-th iteration, let $\nabla L_\eta(C^{(k)})$ be the gradient of the loss function at $C^{(k)}$:

$$\begin{aligned}\nabla L_\eta(C^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{q} Z^{ij}[(x_i^\top c_j^{(k)} - y_{ij})1(|x_i^\top c_j^{(k)} - y_{ij}| \le \eta) \\ + \eta \operatorname{sign}(x_i^\top c_j^{(k)} - y_{ij})1(|x_i^\top c_j^{(k)} - y_{ij}| > \eta)],\end{aligned} \tag{6}$$

where Z^{ij} is a matrix with the j-th column being x_i and the remaining entries 0. Then, we compute a descent direction from

$$V^{(k+1)} \leftarrow \operatorname*{argmin}_{V\in\mathbb{R}^{p\times q}} \langle \nabla L(C^{(k)}), V\rangle \text{ subject to } \|V\|_* \le \lambda,$$

with the solution

$$V^{(k+1)} = -\lambda\, u_1 v_1^\top,$$

where u_1 and v_1 are the leading left and right singular vectors of $\nabla L_\eta(C^{(k)})$.

The algorithm follows Algorithm 1, with different input data and the gradient matrix specified by (6).

2.3 Other robust loss functions

Our framework for developing efficient computation algorithms can easily accommodate a broad class of robust loss functions that are not smooth. Examples of the loss functions are the ℓ_1-loss (the least absolute deviation loss), the check-loss, Tukey's biweight loss, and more; see Hampel et al. (2011).

A scheme is developed as follows. The only necessary adjustment as in Algorithm 1 is calculating the gradient of loss function $\nabla L(\cdot)$. Then, the general updating step is

$$\Theta^{(k+1)} = \Theta^{(k)} + \alpha_{k+1}\left(V^{(k+1)} - \Theta^{(k)}\right),$$

where α_{k+1} is some pre-specified step-size, $V^{(k+1)} = -\lambda\, u_1 v_1^\top$, with u_1 and v_1 being the first left and right singular vectors of $\nabla L(\Theta^{(k)})$.

Table 1 presents gradients for several common loss functions in the context of matrix completion and reduced-rank regression.

Table 1 Gradients under different loss functions for matrix completion ($\nabla L(M)$) and reduced-rank regression ($\nabla L(C)$), $d_{ij} = X_{ij} - M_{ij}$ or $y_{ij} - x_i^\top c_j$ depending on the context

Loss	$\nabla L(\cdot)$
ℓ_1-loss	$\nabla L(M) = \frac{1}{2n}\sum_{t=1}^{n} J_t \operatorname{sign}(d_{i_t j_t})$
$\ell(x) = \lvert x\rvert$	$\nabla L(C) = -\sum_{i=1}^{n}\sum_{j=1}^{q} Z^{ij}\operatorname{sign}(d_{ij})$
Check loss	$\nabla L(M) = \frac{1}{2n}\sum_{t=1}^{n} J_t(1_{(d_{i_t j_t}<0)} - c)$
$\ell_c(x) = x(c - 1_{(x<0)})$	$\nabla L(C) = -\sum_{i=1}^{n}\sum_{j=1}^{q} Z^{ij}(1_{(d_{ij}<0)} - c)$
Tukey's biweight loss	$\nabla L(M) = \frac{1}{2n}\sum_{t=1}^{n} J_t d_{i_t j_t}[1 - (\frac{d_{i_t j_t}}{t})^2]^2 1_{(\lvert d_{i_t j_t}\rvert \le t)}$
$\ell_t(x) = \begin{cases} \frac{t^2}{6} & \text{if } \lvert x\rvert \ge t \\ \frac{t^2}{6}(1 - [1 - (x/t)^2]^3) & \text{o.w.} \end{cases}$	$\nabla L(C) = -\sum_{i=1}^{n}\sum_{j=1}^{q} Z^{ij} d_{ij}[1 - (\frac{d_{ij}}{t})^2]^2 1_{(\lvert d_{ij}\rvert \le t)}$

3 Theory

3.1 Convergence of the algorithms

For self-completeness, we present the theoretical guarantees for the Frank–Wolfe algorithm in the context of robust matrix estimations, together with a simple way to choose the step-sizes.

We prove that by choosing the stepsize properly, the objective functions by using the Huber loss in both matrix completion and reduced-rank regression problems converge to the optimums at the rate of $\mathcal{O}(1/k)$, where k is the iteration counter. The next proposition is for reduced-rank regression problems, and the result for the matrix completion problem can be proved similarly.

Proposition 1 *Consider the loss function* $L_\eta(\cdot) : \mathbb{R}^{n\times p} \to \mathbb{R}$ *constructed from the Huber loss function* (2) *with parameter* η. *For the reduced-rank regression problem* (5), *by the Frank–Wolfe Algorithm with stepsize set as*

$$\alpha_{k+1} = \min\left\{\frac{\nabla L_\eta(C^{(k)})^\top (C^{(k)} - V^{(k+1)})}{L_z \|C^{(k)} - V^{(k+1)}\|^2}, 1\right\}, \text{ for all } k \ge 1,$$

where L_z *is some positive number. Suppose the diameter of the feasible set is* $D := \max_{V_1, V_2 \in S} \|V_1 - V_2\|_F$, *where* $S = \{V : \|V\|_* \le \lambda\}$. *Then, we have that* $L_\eta(C^{(k)})$ *is monotonely decreasing in k, and we have*

$$L_\eta(C^{(k)}) - L_\eta(C^*) \le \frac{2L_z D^2}{k}.$$

Proof Since the Huber loss function is differentiable everywhere, and we have that $\nabla L_\eta(\alpha)$ is Lipschitz-continuous. Thus, with L_z defined above its Lipschitz constant, by Theorem 1 of Freund et al. (2017), we have that the result holds as desired. □

We point out that for the matrix completion problem (1), the result holds by the same argument by letting $L_z = 1$.

Meanwhile, our broad interests include some non-convex losses such as the Tukey's biweight loss. A strategy for handling them is the approximation by a Lipschitz continuous

function with arbitrary precision where simple smoothing techniques are applicable. Upon applying the same stepsizes as discussed above, we can show that the algorithm converges to a stationary point at the same rate; see the analysis of a recent work of Reddi et (2016).

Recently, Charisopoulos et al. (2021) studied the low-rank matrix recovery algorithms with the non-convex rank constraint and non-smooth loss functions. They established optimization convergence rates for a prox-linear method and a subgradient method for matrix completion. They proved that with a sufficient number of observations and an appropriate initialization, both methods are guaranteed to converge to the truth. The prox-linear method possesses a much faster convergence rate ($O(1/2^k)$) but with a higher computational cost at each iteration in solving a convex subproblem. While the subgradient method has a lower cost at each iteration with a subgradient evaluation step and a project step onto the desired region, it has a slower rate. Compared with their algorithms, our method has a lower computational burden in each iteration with no projection required and a relatively slower convergence rate. It is worth studying minimizing a robust loss function directly with the non-convex constraint in the future.

3.2 Statistical properties

We investigate the non-asymptotic error bounds in this section. We first introduce two conditions for both matrix completion and reduced-rank regression models.

Assumption 1 The truth M^* and C^* has rank at most r, $0 < r < \min(p, q)$.

Assumption 2 The noises ξ's are i.i.d. with zero mean and a distribution function F_ξ satisfying

$$F_\xi(m+\eta) - F_\xi(m-\eta) \geq \frac{1}{c_1^2},$$

for any $|m| \leq \eta$ and $\eta > 0$, where $c_1 = c_1(\eta)$ is a constant depending only on η.

Assumption 2 is key on the distribution of the noises.

It is very mild by only requiring non-vanishing probability mass of ξ between $m-\eta$ and $m+\eta$ for a positive η and $|m| \leq \eta$ avoiding assuming instead explicit conditions on its tail probability and/or existence of its moments up to some order.

Since the condition holds for $\eta > 0$ as long as the probability mass of ξ near 0 is not too small, it is easily satisfied by a wide range of distributions including heavy-tailed ones; see more discussion about this assumption and examples in Appendix 1.

3.2.1 Matrix completion

For any matrix A and some linear subspace $\mathcal{M} \subset \mathbb{R}^{p\times q}$, we define $A_\mathcal{M}$ as the projection of A onto $\mathcal{M}$. We consider without loss of generality that $p \geq q > 1$. Recall that A_t $(t = 1, \ldots, n)$ is a $p \times q$ random matrix, independent of $A_{i_t j_t}$ and ξ_t, with one randomly chosen entry $A_{i_t j_t}$ being 1 and the others 0. $M_{i_t j_t}$ can be written as

$$M_{i_t j_t} = \mathrm{tr}(J_t^\top M) = \sum_{i=1}^{p}\sum_{j=1}^{q} J_{t,ij} M_{ij},$$

for all $(i_t, j_t) \in \Omega$ As a working model, we treat Ω_t as uniformly distributed over its support. That is, the probability of $M_{i_t j_t}$ being the t-th observation is $(pq)^{-1}$. This assumes that the observed entries in the target matrix are uniformly sampled at random (Koltchinskii et al., 2011; Rohde & Tsybakov, 2011; Elsener & van de Geer, 2018), and we refer to Klopp (2014) for more discussions. Recht (2011) analyzed the matrix completion model under this assumption. As pointed out in Recht (2011), this is a sampling with replacement scheme and therefore may appear less realistic as it may result in duplicated entries; however, it has the benefit of simplifying the technical proof and assumptions. Overall, it is a reasonable and informative showcase without requiring any prior information on the sampling scheme. If additional information is available in the sampling process, other models such as the weighted sampling model (Negahban & Wainwright, 2012) can be applied.

We first show that the estimator belongs to a restricted set. We consider the singular value decomposition

$$M^* = U \Lambda V^\top,$$

where U is a $p \times q$ matrix, Λ is a $q \times q$ diagonal matrix with diagonal entries the ordered singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0$ and V is a $q \times q$ matrix. For $k = 1, 2, \ldots, q$, let u_k be the k-th column of U, and v_k the k-th column of V. For any positive integer $r \leq \min\{p, q\}$, let U_r be the subspace of $\mathbb{R}^{p \times q}$ spanned by $u_1, \ldots, u_r$, and V_r be the subspace spanned by $v_1, \ldots, v_r$. Define a pair of subspaces in $\mathbb{R}^{p \times q}$ as

$$\mathbb{M}_r(U, V) := \{M \in \mathbb{R}^{p \times q} \mid \mathrm{row}(M) \subseteq V_r,\ \mathrm{col}(M) \subseteq U_r\},$$
$$\overline{\mathbb{M}}_r^{\perp}(U, V) := \{M \in \mathbb{R}^{p \times q} \mid \mathrm{row}(M) \subseteq V_r^{\perp},\ \mathrm{col}(M) \subseteq U_r^{\perp}\},$$

where $\mathrm{row}(M)$ and $\mathrm{col}(M)$ denote the row and column space of M. For simplicity notation, we use $\mathbb{M}_r = \mathbb{M}_r(U, V)$ and $\overline{\mathbb{M}}_r^{\perp} = \overline{\mathbb{M}}_r^{\perp}(U, V)$. Lemma 1 indicates that the estimator $\hat{M}$ belongs to the set

$$\mathbb{M}_0 = \{M \in \mathbb{R}^{p \times q} : ||\Delta_{\overline{\mathbb{M}}_r^{\perp}}||_* \leq 3||\Delta_{\overline{\mathbb{M}}_r}||_* + 4\sum_{k=r+1}^{q} \sigma_k,\ \Delta = M - M^*\}.$$

To establish the error bounds, we need the following technical assumption.

Assumption 3 For any $M \in \mathbb{M}_0$, there exists a real number $L > 1$, such that

$$||M - M^*||_{\max} \leq \frac{L}{\sqrt{pq}} ||M - M^*||_F.$$

Assumptions of this type—referred to as the 'spikiness condition'—are assumed in existing literature on analogous problems, e.g., in Negahban and Wainwright (2012) for matrix completion problems; see also a recent work Fan et al. (2021). Intuitively, this assumption requires that for $M \in \mathbb{M}_0$, the entries of $M - M^*$ are not overly 'spiky', or in other words, relatively evenly distributed; so that the maximum discrepancy is not extremely far away from the averaged discrepancy. We remark that here the term $\frac{1}{\sqrt{pq}}$ relates to the aforementioned uniform sampling scheme setting, under which each entry is observed with the probability $\frac{1}{pq}$

Hence, it reflects an increasingly more difficult high-dimensional problem due to sparse entries in a single copy of large matrix. Instead, if the probability of each entry being observed is a constant independent of p, q, this assumption is not required.

We consider the Lagrangian form of the problem (1):

$$\hat{M} = \underset{M \in \mathbb{R}^{p \times q}}{\operatorname{argmin}} \{ L_\eta(M) + \gamma \|M\|_* \}, \tag{7}$$

where $\gamma > 0$ is the corresponding regularization tuning parameter. Let $\hat{\Delta} = \hat{M} - M^*$ and $\Delta = M - M^*$. Theorem 1 establishes a non-asymptotic upper bound for the error for estimating a M^* of low rank.

Theorem 1 *For problem* (7)*, suppose that Assumption* 1*,* 2*, and* 3 *hold and the noises* ξ_{ij}*'s are distributed symmetrically about zero. Let* $\hat{M}$ *be the solution to problem* (7) *with*

$$\gamma = 2\eta \left\{ 4c_0 \left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)\log(p+q)}{n}} \right] + \sqrt{\frac{2\log(p+q)}{npq}} + \frac{8\log(p+q)}{3n} \right\},$$

with a constant $c_0 > 0$*. When* $n > C(L)\, c_1^2 pr \log(p+q)\log(q+1)$*,*

$$\sqrt{\frac{1}{pq}} \|\hat{\Delta}\|_F \le C_1 c_1^2 \eta \sqrt{\frac{p\log(p+q)\log(q+1)}{n}} \sqrt{2rc_2 + c_3},$$

with probability at least $1 - 3(p+q)^{-1}$*, for some constants* C_1, c_2 *and* c_3 *independent of* n*,* p*, and* q*, and* $C(L)$ *a constant only depending on* L*.*

Theorem 1 is non-asymptotic. γ is chosen based on Lemma 7 in Appendix 2 as twice the upper bound of $\alpha_1(\nabla L_\eta(M^*))$. In Theorem 1, we only require the error terms satisfy Assumption 2, which is $F_\xi(m+\eta) - F_\xi(m-\eta) > \frac{1}{c_1^2}$, for $\eta > 0$ being the parameter in the Huber loss (2) and $|m| < \eta$. Since this assumption is easily satisfied by many heavy-tailed distributions, this result demonstrates the robustness of our method.

We note that in general γ can be

$$K_1 \cdot 2\eta \left\{ 4c_0 \left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)\log(p+q)}{n}} \right] + \sqrt{\frac{2\log(p+q)}{npq}} + \frac{8\log(p+q)}{3n} \right\},$$

for any constant $K_1 \ge 1$. Under the conditions in Theorem 1, we can also derive the upper bounds of the estimation error in nuclear norm based on (23) in the Appendix:

$$\sqrt{\frac{1}{pq}} \|\hat{\Delta}\|_* \le 4C_1 c_1^2 \eta \sqrt{2r} \sqrt{\frac{p\log(p+q)\log(q+1)}{n}} \sqrt{2rc_2 + c_3}.$$

We may discuss the asymptotic properties of $\hat{M}$ when $n \to \infty$. Matrix completion is a hard problem attempting to recover a matrix-valued model parameter with a single incomplete copy from the data generating process. The average estimation error converges to zero in probability as $n \to \infty$. That is, when $rp\log(p+q)\log(q+1) = o(n)$, $(pq)^{-1}\|\hat{\Delta}\|_F^2 \to 0$.

Intuitively, if the rank of M^* is r, then the number of free parameters is at the order of rp. Hence it's reasonable to require a sample size at least of some larger order of rp, so as to recover the model parameters consistently.

Without requiring the Gaussian assumption, our error rate is still comparable to the statistical optimum established by Koltchinskii et al. (2011) for matrix completion problems under a low-rank constraint with Gaussian noises. Compared with the rate in the lower bound given in Theorem 6 of Koltchinskii et al. (2011), our upper bound in Theorem 1 differs only in an additional logarithm term $\sqrt{\log(p+q)\log(q+1)}$ and the η in the Huber loss.

The assumption in Theorem 1 that the model error is symmetrically distributed around 0 is needed in obtaining the upper bound of $\sigma_1(\nabla \mathrm{L}(M^*))$; see the proof of Lemma 7. It assures that $\sigma_1(\mathbb{E}[\nabla \mathrm{L}(M^*)]) = 0$. Similar assumptions are also found in Loh (2017). Thanks to the symmetrization assumption, the convergence can be established with no strong extra requirement on η. Without the symmetrization, as shown in Lemma 7 in the Supplement Material, other conditions are required to control

$$\mathbb{E}\left[\frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}\right] = \int_{M^*_{ij}-\eta}^{M^*_{ij}+\eta} F(X_{ij})dX_{ij} - \eta$$

so that

$$\begin{aligned}\sigma_1(\mathbb{E}[\nabla \mathrm{L}(M^*)]) &= \sigma_1\Big(\frac{1}{2n}\sum_{t=1}^{n}\sum_{i=1}^{p}\sum_{j=1}^{q}\mathbb{E}[J_t J_{t,ij}]\mathbb{E}\left[\frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}\right]\Big) \\ &= \frac{1}{2pq}\sigma_1\Big(\mathbb{E}\left[\frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}\right]_{p\times q}\Big)\end{aligned}$$

is stochastically small enough. With this extra term, the upper bound in Theorem 1 becomes

$$\begin{aligned}\frac{1}{\sqrt{pq}}\|\hat{\Delta}\|_F \le{}& \frac{\text{Constant}\cdot c_1^2\sqrt{2r}}{\sqrt{pq}}\sigma_1\left(\left(\mathbb{E}\left[\frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}\right]\right)_{p\times q}\right) \\ &+ C_1 c_1^2\eta\sqrt{\frac{p\log(p+q)\log(q+1)}{n}}\sqrt{2rc_2 + c_3}\,.\end{aligned} \tag{8}$$

The extra term in (8) may then be viewed as a price paid to achieve robustness against noises with heavy-tailed distributions. This is an impact from applying the robust Huber loss. It is a remarkable different feature from the study on matrix completion with ℓ_2 loss. Nevertheless, it is worth noting that for ℓ_2-loss related studies, conditions are commonly assumed to control the tail probability behavior of the model errors, for example, by the sub-Gaussian distributions. In contrast, our development does not require such assumptions on the tail probability properties, which is the gain in return by applying the Huber loss.

3.2.2 Reduced-rank regression

The problem (5) is also expressed in the Lagrangian form:

$$\hat{C} = \underset{C \in \mathbb{R}^{p \times q}}{\operatorname{argmin}} \{ L_\eta(C) + \gamma \|C\|_* \}, \tag{9}$$

where $\gamma > 0$ is a regularization parameter, and $L_\eta(C)$ is defined in Equation (5).

Again, we point out that the estimator belongs to a restricted set. By applying the singular value decomposition to C^*, we have

$$C^* = U \Lambda V^\top,$$

where $\Lambda = \operatorname{diag}(\sigma_1, \ldots, \sigma_q)$ is the diagonal matrix containing all singular values of C^*. For $r \le \min\{p, q\}$, we define a pair of subspaces of $\mathbb{R}^{p \times q}$ as

$$\mathcal{C}_r(U, V) = \{ M \in \mathbb{R}^{p \times q} \mid \operatorname{row}(M) \subseteq V_r, \operatorname{col}(M) \subseteq U_r \},$$
$$\bar{\mathcal{C}}_r^{\perp}(U, V) = \{ M \in \mathbb{R}^{p \times q} \mid \operatorname{row}(M) \subseteq V_r^{\perp}, \operatorname{col}(M) \subseteq U_r^{\perp} \},$$

where U_r is a subspace spanned by the first r columns of U, and V_r is the subspace spanned by the first r columns of V. For simplicity in notations, we denote by $\mathcal{C}_r = \mathcal{C}_r(U, V)$ and $\bar{\mathcal{C}}_r^{\perp} = \bar{\mathcal{C}}_r^{\perp}(U, V)$. Note that $\mathcal{C}_r$ and $\bar{\mathcal{C}}_r$ are not equal. Lemma 4 indicates that the estimator $\hat{C}$ belongs to the set

$$\mathcal{C}_0 = \left\{ C \in \mathbb{R}^{p \times q} : \|\Delta_{\bar{\mathcal{C}}_r^{\perp}}\|_* \le 3 \|\Delta_{\bar{\mathcal{C}}_r}\|_* + 4 \sum_{k=r+1} \sigma_k, \Delta = C - C^* \right\}.$$

We assume the following conditions on the random design matrix X.

Assumption 4 $x_1, x_2, \ldots, x_n$ are i.i.d. random vectors sampling from a multivariate normal distribution $N(0, \Sigma)$ and without loss of generality, are standardized such that $\|\Sigma\|_F \le 1$. $\sigma_1(\Sigma) \ge \sigma_n(\Sigma) > 0$, where $\sigma_1(\Sigma)$ and $\sigma_n(\Sigma)$ denote the largest and smallest eigenvalues of Σ, respectively.

The multivariate normal distribution and its analogies are commonly assumed in the literature (e.g., Negahban & Wainwright, 2011; Sun et al., 2020; Fan et al., 2021). The setting with Assumption 4 facilitates achieving

the optimal convergence rate; other types of conditions are possible, at the expense of a slower convergence rate.

Theorem 2 establishes a non-asymptotic upper bound for $\|\hat{C} - C^*\|_F$.

Theorem 2 *For problem* (9)*, suppose that Assumption 1 and 2 hold and the noises* ξ_{ij}*'s are distributed symmetrically about zero. Suppose X satisfies Assumption 4. Let* $\hat{C}$ *be the solution to the optimization problem* (9) *with*

$$\gamma = 8\eta\sigma_1(\Sigma) \left(\sqrt{\frac{6n(p+q)}{}} + 3(p+q) \right). \tag{10}$$

Then for $n > C_2 \frac{\sigma_1(\Sigma)}{\sigma_n(\Sigma)} c_1^2 r(p+q)$ *with probability at least* $1 - 3e^{-(p+q)}$,

$$\|\hat{\Delta}\|_F \le C_3 c_1^2 \sqrt{2r}\eta \frac{\sigma_1(\Sigma)}{\sigma_n(\Sigma)} \left(\sqrt{\frac{6(p+q)}{n}} + \frac{3(p+q)}{n} \right),$$

where C_2 *and* C_3 *are constants*.

The value for γ is selected based on Lemma 8 in Appendix 3 as twice the upper bound for $\sigma_1(\nabla L_\eta(C^*))$ according to condition (10). Generally, for any $K_2 \geq 1$ and

$$\gamma = K_2 \cdot 8\eta\sigma_1(\Sigma)\left(\sqrt{6n(p+q)} + 3(p+q)\right),$$

our result remains valid and only differs in constant terms.

Under the same condition, we can establish the error bound in terms of the nuclear norm

$$\|\Delta\|_* \leq 8C_3c_1^2r\eta\frac{\sigma_1(\Sigma)}{\sigma_n(\Sigma)}\left(\sqrt{\frac{6(p+q)}{n}} + \frac{3(p+q)}{n}\right).$$

When $r(p+q) = o(n)$, the Frobenius norm of the error $\|\hat{\Delta}\|_F^2 \to 0$ in probability. Similarly, the robustness of the method is seen as only a mild distributional Assumption 2 is required: $F_\xi(m+\eta) - F_\xi(m-\eta) > \frac{1}{c_1^2}$ for $|m| < \eta$ and $\eta > 0$. Our estimator achieves a comparable convergence rate as that in Negahban and Wainwright (2011) and Rohde and Tsybakov (2011), with the notable difference due to the ℓ_1 in the Huber loss. Meanwhile, our method does not require the errors to follow normal distributions, which is the case in those studies with the ℓ_2 loss. Here assuming symmetricity plays the same role as that in Theorem 1. Based on the same discussions after Theorem 1, if the noises are not symmetrically distributed, then there will be an extra term in the upper bound.

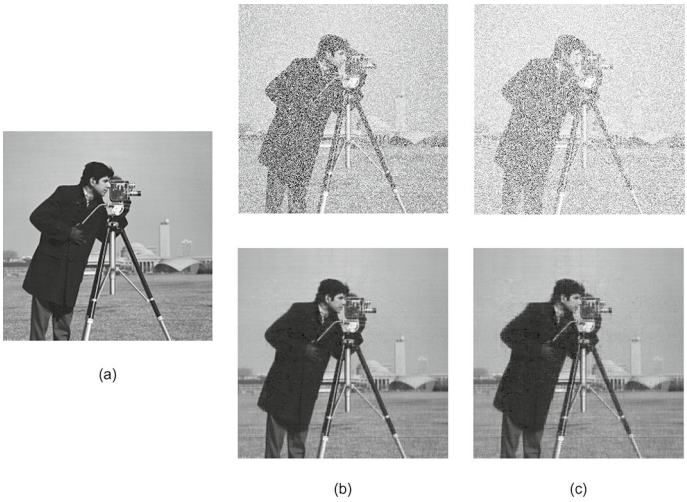
4 Numerical examples

In this section, we conduct an extensive numerical investigation of the proposed method using both simulated and real data sets. In all cases, we choose the tuning parameters by ten-fold cross-validation. Specifically, for matrix completion problems, we first randomly select 90% of the observed entries as training samples and test the results using the remaining 10% samples. We repeat the procedure 10 times and choose the best tuning parameter. With extensive studies on simulated and real data sets, our results provide strong empirical evidence that the proposed method provides robustness under different settings.

4.1 Jester joke data

We first test our method using the Jester joke data set. This data set contains more than 4.1 million ratings for 100 jokes from 73,421 users. This data set is publicly available through http://www.ieor.berkeley.edu/~goldberg/jester-data/. The whole data set contains three sub-datasets, which are: (1) jester-1: 24,983 users who rate 36 or more jokes; (2) jester-2: 23,500 users who rate 36 or more jokes; (3) jester-3: 24,938 users who rate between 15 and 35 jokes. More detailed descriptions can be found in Toh and Yun (2010) and Chen et al. (2012), where the authors consider the nuclear-norm based approach to conduct matrix completion.

Due to the large number of users, we randomly select $n_u = 1000$ users' ratings from the datasets. Since many entries are unknown, we cannot compute the relative error using every entry. Instead, we take the metric of the normalized mean absolute error (NMAE) to measure the accuracy of the estimator $\hat{M}$:



(a) (b) (c)

Fig. 1 **a** We test our method on a 512 × 512 Cameraman image. **b** A sample noisy image with heavy-tailed noises and 40% missing entries. **c** A sample noisy image with heavy-tailed noises and 60% missing entries

$$\text{NMAE} = \frac{\sum_{(j,k)\in\Omega} |\hat{M}_{jk} - M^0_{jk}|}{|\Omega|(r_{\max} - r_{\min})},$$

where $r_{\min}$ and $r_{\max}$ denote the lower and upper bounds for the ratings, respectively. In the Jester joke data set, the range is $[-10, 10]$. Thus, we have $r_{\max} - r_{\min} = 20$.

In each iteration, we first randomly select n_u users, and then randomly permute the ratings from the users to generate $M^0 \in \mathbb{R}^{n_u \times 100}$. Next, we uniformly sample SR for SR $\in \{15\%, 20\%, 25\%\}$ entries to generate a set of observed indices Ω. Note that we can only observe the entry (j, k) if $(j, k) \in \Omega$, and $M^0_{j,k}$ is available. Thus, the actual sampling ratio is less than the input SR. We consider different settings of n_u and SR, and we report the averaged NMAE and running times in Table 2 after running each setting 100 times. We compare robust methods with ℓ_1 loss, Huber loss, and Tukey loss with the non-robust ℓ_2 loss. From Table 2, we see that robust matrix completion methods work promisingly.

4.2 Cameraman image denoising

We test our method using the popular Cameraman image, which is widely used in image processing literature. We consider the "Cameraman" image with 512×512 pixels as shown in Fig. 1a. We then generate random noise by first adding independent Gaussian noise to each pixel with a standard deviation set as 3. Then, we add some heavy-tailed noises by randomly choosing 10% pixels and replace the coefficient as 100 or 0. Furthermore, we randomly select 40% or 60% pixels as missing entries. We provide two typical simulated noisy images in the above of Fig. 1b, c, and provide the recovered images using the Tukey approach below them. The recovered images provide visual evidence that our method is robust to heavy-tailed noises in practice. In addition, in Table 3, we provide the averaged NMAE with standard deviations of different approaches after repeating the data generating schemes 100 times. For the effective picture recovery and the NMAE, we

Table 2 Averaged normalized mean absolute error with standard deviations in the parentheses for different methods using Jester joke data set under different data generating schemes after 100 runs

Example	$(n_u$, SR)	Huber	Tukey	ℓ_2	ℓ_1
jester-1	(1000, 0.15)	0.155(0.006)	0.154(0.007)	0.173(0.009)	0.168(0.006)
	(1000, 0.20)	0.152(0.005)	0.150(0.006)	0.171(0.009)	0.165(0.006)
	(1000, 0.25)	0.145(0.005)	0.143(0.006)	0.168(0.008)	0.153(0.004)
	(1500, 0.15)	0.159(0.006)	0.155(0.006)	0.177(0.007)	0.169(0.009)
	(1500, 0.20)	0.154(0.006)	0.152(0.006)	0.174(0.007)	0.166(0.008)
	(1500, 0.25)	0.151(0.006)	0.150(0.007)	0.173(0.006)	0.162(0.008)
	(2000, 0.15)	0.160(0.006)	0.159(0.006)	0.180(0.006)	0.169(0.006)
	(2000, 0.20)	0.158(0.007)	0.156(0.006)	0.178(0.006)	0.164(0.007)
	(2000, 0.25)	0.155(0.005)	0.154(0.006)	0.175(0.006)	0.161(0.006)
jester-2	(1000, 0.15)	0.163(0.007)	0.161(0.008)	0.176(0.007)	0.169(0.008)
	(1000, 0.20)	0.160(0.007)	0.159(0.008)	0.172(0.007)	0.167(0.008)
	(1000, 0.25)	0.158(0.006)	0.155(0.007)	0.170(0.008)	0.166(0.007)
	(1500, 0.15)	0.166(0.007)	0.164(0.008)	0.178(0.008)	0.171(0.007)
	(1500, 0.20)	0.164(0.006)	0.161(0.007)	0.176(0.007)	0.168(0.007)
	(1500, 0.25)	0.161(0.007)	0.160(0.007)	0.173(0.007)	0.164(0.008)
	(2000, 0.15)	0.170(0.006)	0.168(0.008)	0.180(0.007)	0.173(0.007)
	(2000, 0.20)	0.166(0.007)	0.165(0.008)	0.177(0.007)	0.171(0.008)
	(2000, 0.25)	0.163(0.006)	0.163(0.008)	0.175(0.008)	0.169(0.007)
jester-3	(1000, 0.15)	0.175(0.008)	0.173(0.008)	0.184(0.008)	0.179(0.008)
	(1000, 0.20)	0.173(0.008)	0.171(0.008)	0.181(0.007)	0.177(0.008)
	(1000, 0.25)	0.170(0.008)	0.168(0.009)	0.179(0.008)	0.176(0.008)
	(1500, 0.15)	0.177(0.008)	0.176(0.008)	0.187(0.008)	0.181(0.008)
	(1500, 0.20)	0.174(0.007)	0.174(0.008)	0.185(0.009)	0.178(0.009)
	(1500, 0.25)	0.173(0.008)	0.172(0.008)	0.184(0.008)	0.176(0.008)
	(2000, 0.15)	0.179(0.008)	0.178(0.008)	0.188(0.008)	0.182(0.008)
	(2000, 0.20)	0.177(0.009)	0.175(0.008)	0.187(0.009)	0.180(0.008)
	(2000, 0.25)	0.174(0.008)	0.172(0.008)	0.185(0.008)	0.177(0.007)

Table 3 Averaged normalized mean absolute error with standard deviations in the parentheses for different methods using Lena image after 100 runs

Missing rate	Huber	Tukey	ℓ_2	ℓ_1
40%	0.067(0.004)	0.062(0.005)	0.083(0.008)	0.079(0.006)
50%	0.071(0.005)	0.065(0.006)	0.089(0.011)	0.084(0.007)
60%	0.074(0.005)	0.069(0.007)	0.092(0.015)	0.088(0.007)

conclude that robust matrix completion has promising performance with partial and noisy observations.

4.3 Simulations

We first consider several similar simulation settings as described in She and Chen (2017) to compare our method with their robust reduced-rank regression (RR4) method. In all cases, we focus on testing the robustness by artificially introducing data corruption and outliers.

Setting 1: We first consider a low-dimensional case where $n = 100$, $p = 12$, $q = 8$ and $r = 3$ or 5. We construct the design matrix X by generating its n rows by independently sampling from $N(0, \Sigma_0)$, where we consider highly correlated covariates by letting the diagonal elements of Σ_0 be 1's and setting its off-diagonal elements as 0.5. For the noise matrix Ξ, we sample each row of Ξ independently from $N(0, \sigma^2\Sigma_1)$, where Σ_1 is the q-dimensional identity matrix, and σ is set as 1. Next, we construct the coefficient matrix C^*. We generate $C^* = B_1 B_2^\top$, where $B_1 \in \mathbb{R}^{p\times r}$, $B_2 \in \mathbb{R}^{q\times r}$, and all entries of B_1 and B_2 are independently sampled from $N(0, 1)$. We then add outliers with a matrix U^* by setting the first $\alpha\%$ n rows of U^* as nonzero, where $\alpha \in \{30, 35, \ldots, 50\}$ is the proportion of outliers, and the j-th entry of any outlier row of U^* is the product of a Rademacher random variable and a scalar $\alpha \in \{0.75, 1\}$ times the sample standard deviation of the j-th column of XC^*. Finally, we set the response matrix $Y = XC^* + U^* + \Xi$. We report the mean and standard deviation of the mean squared error (MSE) from 200 runs, where

$$\text{MSE}(\hat{C}) = \|XC^* - X\hat{C}\|_F^2/(qn).$$

In addition, we also report the mean and standard deviation of the mean squared estimation error, where

$$\text{MSE}(\hat{C}) = \|\hat{C} - C^*\|_F^2/(qp).$$

Setting 2: We then test our method on heavy-tailed noise. Same as Setting 1, we let $n = 100, p = 12, q = 8$, and $r = 2$, 3, or 4, and consider the same generating scheme to construct the design matrix X, and then generate the noise matrix by the heavy-tailed t-distribution with a degree of freedom 3 or 5. Furthermore, we add outliers by the same generating scheme as in Setting 1 to generate U^* and letting α = 0.5, 0.75 or 1.

Setting 3: We consider a high-dimensional setting where $n = 100, p = 50$ and $q = 50$, and $r = 3$ or 5, where there are 2, 500 > 100 parameters in the matrix C to be estimated. We consider the same data generating scheme as in Setting 1.

Setting 4: Finally, we consider an ultrahigh-dimensional setting where $n = 300$, $p = 100$ and $q = 400$, and $r = 3$ or 5, where there are $40,000 \gg 300$ parameters to be estimated. We consider the same data generating scheme as in Setting 1.

The results are shown in Tables 4, 5, 6, and 7. We compare our method incorporating Huber and Tukey loss functions with the RR4 method when it is applicable. We note that for high-dimensional Settings 3 and 4, the RR4 method of She and Chen (2017) cannot be applied here because one of the iterations in their algorithm is not defined. We compare our method with another robust method where we use ℓ_1 loss in place of the Huber loss in the objective with the nuclear norm constraint (Denoted as ℓ_1). In all four settings, both Huber loss and Tukey loss achieve very promising performance, and Tukey loss slightly outperforms Huber loss in settings with outliers.

Table 4 Sample average of MSE($\hat{C}$) and MSE($\hat{C}$) for Setting 1 under different settings with sample standard deviation in parentheses after 200 runs

r	α	σ%	MSE($\hat{C}$)			MSE($\hat{C}$)		
			Huber	Tukey	R^4	Huber	Tukey	R^4
3	0.75	30%	0.71(0.28)	0.56(0.31)	0.95(0.75)	0.12(0.05)	0.09(0.06)	0.23(0.12)
		35%	0.82(0.44)	0.63(0.38)	1.23(1.09)	0.13(0.07)	0.09(0.06)	0.25(0.17)
		40%	0.96(0.49)	0.83(0.58)	1.46(1.26)	0.16(0.08)	0.13(0.08)	0.28(0.20)
		45%	1.11(0.97)	0.97(0.89)	1.57(1.24)	0.18(0.09)	0.15(0.09)	0.30(0.21)
		50%	1.23(1.01)	1.03(0.95)	1.69(1.31)	0.19(0.11)	0.16(0.10)	0.33(0.23)
	1.00	30%	1.02(0.42)	0.89(0.48)	1.93(1.88)	0.12(0.07)	0.10(0.11)	0.25(0.37)
		35%	1.12(0.46)	0.96(0.51)	2.06(2.01)	0.18(0.08)	0.14(0.12)	0.34(0.34)
		40%	1.34(0.64)	1.20(0.52)	2.59(2.12)	0.22(0.10)	0.20(0.14)	0.42(0.35)
		45%	1.65(0.77)	1.39(0.85)	2.88(2.35)	0.27(0.12)	0.24(0.15)	0.48(0.40)
		50%	1.83(0.84)	1.60(1.05)	3.28(2.76)	0.29(0.13)	0.25(0.18)	0.53(0.45)
5	0.75	30%	0.78(0.34)	0.64(0.44)	1.35(1.03)	0.13(0.05)	0.10(0.07)	0.26(0.15)
		35%	0.87(0.42)	0.72(0.48)	1.78(1.15)	0.14(0.08)	0.11(0.08)	0.29(0.21)
		40%	0.94(0.67)	0.88(0.58)	1.63(1.32)	0.17(0.08)	0.14(0.09)	0.31(0.23)
		45%	1.15(0.82)	0.92(0.95)	1.85(1.49)	0.19(0.10)	0.16(0.10)	0.34(0.26)
		50%	1.32(1.13)	1.20(1.06)	2.04(1.63)	0.21(0.13)	0.19(0.11)	0.39(0.31)
	1.00	30%	0.71(0.62)	0.65(0.44)	2.02(1.95)	0.13(0.08)	0.10(0.12)	0.31(0.34)
		35%	1.19(0.54)	0.99(0.63)	2.13(2.15)	0.19(0.09)	0.17(0.13)	0.39(0.39)
		40%	1.45(0.71)	1.15(0.75)	2.64(2.27)	0.23(0.11)	0.23(0.15)	0.45(0.48)
		45%	1.77(0.84)	1.44(0.93)	3.06(2.48)	0.28(0.14)	0.26(0.16)	0.53(0.55)
		50%	1.90(0.95)	1.61(1.03)	3.31(2.78)	0.31(0.15)	0.28(0.19)	0.61(0.68)

5 Intermediate theoretical results

Our estimators (1) and (5) are penalized M-estimators. We exploit the framework of Negahban et al. (2012) in studying their statistical properties. Negahban et al. (2012) elaborates the notion of decomposability associated with some penalty function, which is a key property for establishing the restricted strong convexity (RSC) property and the error bounds of the penalized estimators.

For self-completeness, we outline the decomposability of penalizing with the nuclear norm, and then derive the restricted strong convexity property for both models under the Huber loss function.

5.1 Decomposability of nuclear norm

A norm $\|\cdot\|$ is decomposable with respect to a pair of subspace if for all $A \in \mathcal{M}$ and $B \in \overline{\mathcal{M}}^{\perp}$ with $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ a pair of subspace in $\mathbb{R}^{p\times q}$ satisfy

$$\|A + B\| = \|A\| + \|B\|.$$

To illustrate the decomposability of nuclear norm, recall

Table 5 Sample average of MSE($X\hat{C}$) and MSE($\hat{C}$) for Setting 2 under different settings with sample standard deviation in parentheses after 200 runs

d.o.f.	α	r	MSE($X\hat{C}$)			MSE($\hat{C}$)		
			Huber	Tukey	R^4	Huber	Tukey	R^4
3	0.50	2	0.90(0.45)	0.70(0.42)	1.02(1.11)	0.14(0.08)	0.11(0.07)	0.20(0.14)
		3	0.72(0.30)	0.52(0.28)	0.88(0.36)	0.12(0.05)	0.09(0.04)	0.15(0.17)
		4	0.98(0.88)	0.65(0.81)	1.24(1.03)	0.17(0.14)	0.12(0.13)	0.39(0.28)
	0.75	2	0.59(0.27)	0.40(0.23)	0.64(0.33)	0.10(0.05)	0.06(0.04)	0.30(0.21)
		3	0.46(0.24)	0.23(0.14)	0.45(0.17)	0.08(0.04)	0.04(0.02)	0.12(0.09)
		4	1.01(1.15)	0.66(1.06)	1.32(1.03)	0.18(0.19)	0.12(0.18)	0.19(0.12)
	1.00	2	0.36(0.20)	0.20(0.12)	0.46(0.26)	0.06(0.03)	0.03(0.02)	0.08(0.03)
		3	0.36(0.17)	0.15(0.09)	0.49(0.31)	0.07(0.03)	0.03(0.02)	0.09(0.04)
		4	0.84(0.60)	0.50(0.58)	0.93(0.73)	0.15(0.11)	0.10(0.11)	0.20(0.05)
5	0.50	2	0.91(0.48)	0.74(0.45)	1.02(1.11)	0.14(0.08)	0.11(0.07)	0.15(0.06)
		3	0.69(0.36)	0.51(0.36)	1.32(0.48)	0.12(0.06)	0.08(0.06)	0.16(0.04)
		4	0.95(0.85)	0.76(0.91)	1.42(0.58)	0.17(0.15)	0.14(0.16)	0.19(0.09)
	0.75	2	0.51(0.26)	0.36(0.20)	0.49(0.30)	0.08(0.05)	0.06(0.03)	0.08(0.06)
		3	0.44(0.19)	0.21(0.12)	0.66(0.22)	0.08(0.04)	0.04(0.02)	0.13(0.06)
		4	0.68(0.62)	0.63(0.67)	0.71(1.03)	0.18(0.29)	0.12(0.28)	0.23(0.14)
	1.00	2	0.37(0.21)	0.21(0.16)	0.29(0.22)	0.06(0.04)	0.03(0.03)	0.06(0.03)
		3	0.39(0.16)	0.13(0.08)	0.45(0.31)	0.07(0.03)	0.02(0.01)	0.09(0.04)
		4	0.42(0.39)	0.38(0.34)	0.92(0.73)	0.17(0.15)	0.12(0.16)	0.20(0.05)

$$\mathrm{M}_r(U,V) := \{M \in \mathbb{R}^{p\times q} \mid \mathrm{row}(M) \subseteq V_r,\ \mathrm{col}(M) \subseteq U_r\},$$
$$\overline{\mathrm{M}}_r^{\perp}(U,V) := \{M \in \mathbb{R}^{p\times q} \mid \mathrm{row}(M) \subseteq V_r^{\perp},\ \mathrm{col}(M) \subseteq U_r^{\perp}\}.$$

Note that $\mathrm{M}_r \neq \overline{\mathrm{M}}_r$. Since U and V both have orthogonal columns, nuclear norm is decomposable with respect to the pair $(\mathrm{M}_r, \overline{\mathrm{M}}_r^{\perp})$. Note that if the rank of M^* is equal or smaller than r, then U_r and V_r equal to or contain the column and row spaces of M^* respectively, and $M^* \in \mathrm{M}_r(U,V)$.

We present key intermediate results as lemmas below. The proofs of the lemmas are given in the Appendix.

5.2 Results for matrix completion

The decomposability leads to the first lemma, which is a special case of Lemma 1 in Negahban et al. (2012). It provides an upper bound for $\|\hat{\Delta}_{\overline{\mathrm{M}}_r^{\perp}}\|_*$.

Lemma 1 *For any γ satisfying*

$$\gamma \geq 2\sigma_1(\nabla \mathrm{L}_\eta(M^*)),$$

the error $\hat{\Delta}$ satisfies

Table 6 Sample average of MSE($X\hat{C}$) and MSE($\hat{C}$) for Setting 3 under different settings with sample standard deviation in parentheses after 200 runs

r	α	σ%	MSE($X\hat{C}$)			MSE($\hat{C}$)		
			Huber	Tukey	ℓ_1	Huber	Tukey	ℓ_1
3	0.75	30%	1.23(1.19)	1.07(1.08)	1.43(1.02)	0.02(0.02)	0.02(0.02)	0.04(0.02)
		35%	1.44(1.64)	1.28(1.19)	1.65(1.10)	0.03(0.03)	0.02(0.03)	0.05(0.03)
		40%	1.65(1.25)	1.38(1.98)	1.99(1.19)	0.03(0.04)	0.03(0.03)	0.06(0.03)
		45%	1.72(1.33)	1.54(1.41)	2.44(1.34)	0.03(0.04)	0.03(0.03)	0.08(0.04)
		50%	1.83(1.46)	1.61(1.52)	2.07(1.51)	0.04(0.05)	0.03(0.04)	0.08(0.05)
	1.00	30%	1.34(1.89)	1.15(1.63)	1.51(0.88)	0.03(0.02)	0.02(0.02)	0.03(0.04)
		35%	1.45(1.84)	1.27(1.53)	1.64(0.94)	0.03(0.03)	0.02(0.03)	0.04(0.02)
		40%	1.52(1.80)	1.38(1.44)	1.82(1.05)	0.03(0.04)	0.03(0.03)	0.04(0.03)
		45%	1.63(1.95)	1.46(1.67)	2.03(1.12)	0.04(0.04)	0.03(0.04)	0.05(0.03)
		50%	1.75(2.03)	1.53(1.71)	2.19(1.40)	0.04(0.04)	0.03(0.04)	0.08(0.07)
5	0.75	30%	1.31(1.22)	1.12(1.05)	1.46(1.10)	0.02(0.03)	0.02(0.03)	0.04(0.03)
		35%	1.50(1.72)	1.31(1.25)	1.73(1.16)	0.03(0.04)	0.02(0.03)	0.05(0.05)
		40%	1.73(1.36)	1.44(2.05)	2.03(1.31)	0.03(0.05)	0.03(0.04)	0.07(0.06)
		45%	1.81(1.41)	1.63(1.49)	2.58(1.42)	0.04(0.05)	0.03(0.05)	0.09(0.05)
		50%	1.90(1.55)	1.72(1.63)	2.19(1.59)	0.04(0.06)	0.04(0.04)	0.11(0.06)
	1.00	30%	1.39(1.74)	1.23(1.85)	1.67(0.95)	0.03(0.04)	0.03(0.03)	0.05(0.06)
		35%	1.55(1.92)	1.35(1.53)	1.79(1.09)	0.03(0.05)	0.03(0.03)	0.06(0.05)
		40%	1.67(1.79)	1.40(1.61)	1.96(1.21)	0.04(0.04)	0.04(0.04)	0.06(0.04)
		45%	1.74(1.85)	1.58(1.52)	2.14(1.37)	0.05(0.05)	0.04(0.05)	0.07(0.05)
		50%	1.89(1.93)	1.63(1.66)	2.25(1.19)	0.06(0.05)	0.05(0.05)	0.10(0.06)

$$\|\hat{\Delta}_{\overline{M}_r^{\perp}}\|_* \le 3\|\hat{\Delta}_{\overline{M}_r}\|_* + 4\sum_{k=r+1}^{q}\sigma_k.$$

Lemma 1 indicates that the estimator $\hat{M}$ belongs to the set

$$\mathcal{M}_0 = \{M \in \mathbb{R}^{p\times q} : \|\Delta_{\overline{M}_r^{\perp}}\|_* \le 3\|\Delta_{\overline{M}_r}\|_* + 4\sum_{k=r+1}^{q}\sigma_k,\ \Delta = M - M^*\}.$$

Note that if the rank of M^* is no greater than r, then $\sum_{k=r+1}^{q}\sigma_k = 0$ and the projection of the error on $\overline{M}_r^{\perp}$ is solely controlled by the projection of error on $\overline{M}_r$, so as the error itself, since

$$\|\hat{\Delta}\|_* \le \|\hat{\Delta}_{\overline{M}_r^{\perp}}\|_* + \|\hat{\Delta}_{\overline{M}_r}\|_* \le 4\|\hat{\Delta}_{\overline{M}_r}\|_*.$$

Now, consider the quantity

$$\delta L_\eta(M, M^*) = L_\eta(M) - L_\eta(M^*) - \langle \nabla L_\eta(M^*), \Delta\rangle.$$

For simplicity, we sometimes refer to $\delta L_\eta(M, M^*)$ as δL_η. The next Lemma gives a lower bound of $\delta L_\eta(M, M^*)$, which is used to establish restricted strong convexity (RSC) and the

Table 7 Sample average of MSE($X\hat{C}$) and MSE($\hat{C}$) for Setting 4 under different settings with sample standard deviation in parentheses after 200 runs

r	α	σ%	MSE($X\hat{C}$)			MSE($\hat{C}$)		
			Huber	Tukey	ℓ_1	Huber	Tukey	ℓ_1
3	0.75	30%	1.32(1.22)	1.23(1.13)	1.52(1.23)	0.04(0.03)	0.04(0.03)	0.06(0.03)
		35%	1.57(1.52)	1.35(1.25)	1.81(1.34)	0.05(0.04)	0.04(0.03)	0.07(0.04)
		40%	1.65(1.46)	1.53(1.39)	2.03(1.51)	0.05(0.05)	0.04(0.04)	0.08(0.05)
		45%	1.79(1.49)	1.64(1.57)	2.21(1.53)	0.06(0.05)	0.05(0.04)	0.10(0.06)
		50%	1.90(1.53)	1.75(1.63)	2.28(1.65)	0.07(0.06)	0.05(0.05)	0.12(0.08)
	1.00	30%	1.41(2.01)	1.34(1.45)	1.63(1.09)	0.04(0.03)	0.04(0.03)	0.06(0.04)
		35%	1.60(2.15)	1.49(1.61)	1.85(1.30)	0.06(0.05)	0.05(0.04)	0.08(0.05)
		40%	1.69(2.09)	1.60(1.77)	2.09(1.49)	0.06(0.06)	0.05(0.04)	0.10(0.06)
		45%	1.81(2.13)	1.68(1.59)	2.20(1.53)	0.08(0.06)	0.06(0.04)	0.11(0.08)
		50%	1.95(2.33)	1.79(2.12)	2.29(1.47)	0.09(0.06)	0.06(0.05)	0.12(0.10)
5	0.75	30%	1.45(1.39)	1.31(1.33)	1.65(1.19)	0.05(0.03)	0.04(0.03)	0.07(0.03)
		35%	1.61(1.48)	1.45(1.39)	1.89(1.25)	0.06(0.04)	0.05(0.03)	0.09(0.05)
		40%	1.82(1.52)	1.62(1.51)	2.19(1.35)	0.07(0.06)	0.05(0.06)	0.10(0.06)
		45%	1.95(1.36)	1.78(1.59)	2.25(1.39)	0.08(0.06)	0.06(0.06)	0.12(0.07)
		50%	2.04(1.58)	1.85(1.57)	2.37(1.44)	0.08(0.07)	0.07(0.06)	0.13(0.08)
	1.00	30%	1.53(1.65)	1.38(1.42)	1.70(1.08)	0.05(0.03)	0.04(0.04)	0.08(0.04)
		35%	1.66(1.74)	1.49(1.49)	1.83(1.29)	0.06(0.05)	0.05(0.04)	0.10(0.06)
		40%	1.79(1.79)	1.66(1.53)	2.07(1.27)	0.07(0.06)	0.06(0.06)	0.11(0.07)
		45%	1.92(1.70)	1.81(1.62)	2.15(1.43)	0.08(0.07)	0.06(0.06)	0.12(0.08)
		50%	2.08(1.85)	1.93(1.59)	2.34(1.52)	0.09(0.08)	0.07(0.07)	0.14(0.08)

upper bound for the error. The key to proving this lemma includes Lemma 1 and the application of empirical process techniques.

Lemma 2 (*Lower bound of* $\delta L_\eta(M, M^*)$) *Suppose Assumption 1 and 2 hold, and that the regularization parameter in optimization problem* (7) *satisfies*

$$\gamma \geq 2\sigma_1(\nabla L_\eta(M^*)).$$

Then for any $x > 0$ *and* $M \in \{M : \|M - M^*\|_{\max} \leq \eta\} \cap \mathbb{M}_0$,

$$\delta L_\eta(M, M^*) \geq \frac{1}{4c_1^2 pq}\|\Delta\|_F^2 - \{32\sqrt{2r\eta c_0}\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\log(q+1)}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x\eta^2}{npq}} + \frac{8x\eta}{3n}\}\|\Delta\|_F,$$

with probability at least $1 - e^{-x}$.

By controlling the negative term, we have the restricted strong convexity property.

Lemma 3 (*Restricted Strong Convexity*) *Suppose that all the conditions in Lemma 2 and Assumption 3 hold. For* $M \in \{M : \|M - M^*\|_{\max} \le \eta\} \cap \mathbb{M}_0$, *with probability at least* $1 - e^{-(p+q)}$,

$$\delta \mathcal{L}_\eta(M, M^*) \ge \frac{1}{8c_1^2 pq} \|\Delta\|_F^2,$$

for $n > C(L)\, c_1^2 pr \log(p+q) \log(q+1)$, *where* $C(L)$ *is a a constant only depending on* L.

5.3 Results of reduced-rank regression

Recall

$$\mathbb{C}_r(U, V) = \{M \in \mathbb{R}^{p\times q} \mid \mathrm{row}(M) \subseteq V_r,\ \mathrm{col}(M) \subseteq U_r\},$$
$$\bar{\mathbb{C}}_r^{\perp}(U, V) = \{M \in \mathbb{R}^{p\times q} \mid \mathrm{row}(M) \subseteq V_r^{\perp},\ \mathrm{col}(M) \subseteq U_r^{\perp}\}.$$

Lemma 1 can be easily extended to $\hat{C}$

Lemma 4 *For any* γ *satisfying*

$$\gamma \ge 2\sigma_1(\nabla \mathcal{L}_\eta(C^*)),$$

$\hat{\Delta} = \hat{C} - C^*$ *satisfies*

$$\|\hat{\Delta}_{\bar{\mathbb{C}}_r^{\perp}}\|_* \le 3\|\hat{\Delta}_{\bar{\mathbb{C}}_r}\|_* + 4 \sum_{k=r+1} \sigma_k.$$

Lemma 4 indicates that the estimator $\hat{C}$ belongs to the set

$$\mathbb{C}_0 = \{C \in \mathbb{R}^{p\times q} : \|\Delta_{\bar{\mathbb{C}}_r^{\perp}}\|_* \le 3\|\Delta_{\bar{\mathbb{C}}_r}\|_* + 4 \sum_{k=r+1} \sigma_k, \Delta = C - C^*\}.$$

The next result is to establish the RSC condition. Consider the quantity

$$\delta \mathcal{L}_\eta(C, C^*) = \mathcal{L}_\eta(C) - \mathcal{L}_\eta(C^*) - \langle \nabla \mathcal{L}_\eta(C^*), \Delta \rangle.$$

Lemma 5 (*Lower bound of* $\delta \mathcal{L}_\eta(C, C^*)$) *Consider the reduced-rank regression problem* (9). *Suppose that Assumption* 1, 2 *and* 4 *hold, and the noises* ξ_i *are distributed symmetrically about zero. Suppose the regularization parameter in optimization problem* (9) *satisfies*

$$\gamma \ge 2\sigma_1(\nabla \mathcal{L}_\eta(C^*)).$$

Then for any $x > 0$ *and* $C \in \{C : \|C - C^*\|_F \le \eta\} \cap \mathbb{C}_0$,

$$\delta \mathcal{L}_\eta(C, C^*) \ge \frac{n\sigma_n(\Sigma)}{2c_1^2} \|\Delta\|_F^2 - 48\sqrt{2r}\eta\sigma_1(\Sigma)\Big(\sqrt{4n(p+q) + 2nx} + 2(p+q) + x\Big)\|\Delta\|_F,$$

with probability at least $1 - e^{-x}$.

By controlling the negative term and setting the right side to be greater than 0, we have the restricted strong convexity property.

Lemma 6 (*Restricted Strong Convexity*) *Suppose that all the conditions in Lemma 5 hold, then for* $C \in \{C : \|C - C^*\|_F \leq \eta\} \cap \mathbb{C}_0$ *and* $n > C_2 \frac{\sigma_1(\Sigma)}{\sigma_n(\Sigma)} c_1^2 r(p+q)$, *where* C_2 *is a constant*,

$$\delta \mathrm{L}_\eta(C, C^*) \geq \frac{n\sigma_n(\Sigma)}{4c_1^2} \|\Delta\|_F^2,$$

with probability at least $1 - (p+q)^{-1}$.

Appendices

Appendix 1: More on the assumption on the model errors

A key assumption in Theorem 1 and Theorem 2 is that the noises ξ's are i.i.d. with zero mean and a distribution function F_ξ satisfying

$$F_\xi(m+\eta) - F_\xi(m-\eta) > \frac{1}{c_1(\eta)^2}, \tag{11}$$

for any $|m| \leq \eta$ and some $\eta > 0$, where c_1 is a positive constant depending only on η. This is the same as requiring $Pr(\xi \in [m-\eta, m+\eta])$ to be always positive for any $|m| \leq \eta$ and $\eta > 0$. Since $E(\xi) = 0$ and $0 \in [m-\eta, m+\eta]$, this condition holds as long as the probability mass near 0 is not too small, which is easily satisfied by a large class of distributions including heavy-tailed ones. As an example, Fig. 2 gives the distribution of a *t*-distribution with degree of freedom being 3. The area of the grey part represents $F_\xi(m+\eta) - F_\xi(m-\eta)$ when $m = 1$ and $\eta = 2$. Since the density function near 0 is strictly bounded from below, the required condition (11) holds for $\eta > 0$.

The Huber contamination model also satisfies Assumption 2. Specifically, suppose the errors ξ's follow a Huber contamination model $(1-c)F + cG$ with F being the distribution function of a normal random variable. Then $Pr(\xi \in [m-\eta, m+\eta]) = (1-c)\{F(m+\eta) - F(m-\eta)\} + c\{G(m+\eta) - G(m-\eta)\}$. Then the first term creates no issue. Assumption 2 is easily met if G in the second term is a continuous distribution with zero mean. When G is from a discrete distribution, it is a step function. Then the second term is either 0 or a value bounded above from 1. Overall, Assumption 2 is satisfied.

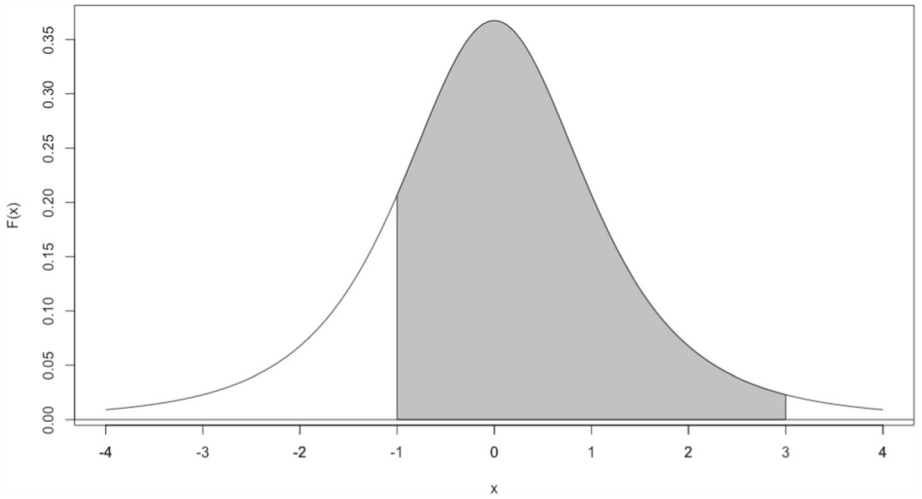


Fig. 2 The distribution of a t-distribution with degree of freedom being 3. The area of the grey part represents $F_\xi(m+\eta) - F_\xi(m-\eta)$ when $m = 1$ and $\eta = 2$

Appendix 2: Proof for matrix completion

This section presents the proof related to the matrix completion.

Proof of Lemma 1 Note that

$$M^*_{\mathbf{M}_r} + M^*_{\overline{\mathbf{M}}_r^\perp} = \sum_{k=1}^{r} u_k \sigma_k v_k^\top + \sum_{k=r+1}^{q} u_k \sigma_k v_k^\top = M^*.$$

Using triangle inequalities and the decomposability of nuclear norm on $\mathbf{M}_r$ and $\overline{\mathbf{M}}_r^\perp$,

$$\begin{aligned} ||\hat{M}||_* &= ||M^* + \hat{\Delta}||_* = ||M^*_{\mathbf{M}_r} + M^*_{\overline{\mathbf{M}}_r^\perp} + \hat{\Delta}_{\mathbf{M}_r} + \hat{\Delta}_{\overline{\mathbf{M}}_r^\perp}||_* \\ &\geq ||M^*_{\mathbf{M}_r} + \hat{\Delta}_{\overline{\mathbf{M}}_r^\perp}||_* - ||M^*_{\overline{\mathbf{M}}_r^\perp} + \hat{\Delta}_{\overline{\mathbf{M}}}||_* \\ &\geq ||M^*_{\mathbf{M}_r}||_* + ||\hat{\Delta}_{\overline{\mathbf{M}}_r^\perp}||_* - ||M^*_{\overline{\mathbf{M}}_r^\perp}||_* - ||\hat{\Delta}_{\overline{\mathbf{M}}}||_*. \end{aligned}$$

Thus,

$$\begin{aligned} ||M^*||_* - ||\hat{M}||_* &\leq ||M^*||_* - ||M^*_{\mathbf{M}_r}||_* - ||\hat{\Delta}_{\overline{\mathbf{M}}_r^\perp}||_* + ||M^*_{\overline{\mathbf{M}}_r^\perp}||_* + ||\hat{\Delta}_{\overline{\mathbf{M}}}||_* \\ &= 2||M^*_{\overline{\mathbf{M}}_r^\perp}||_* + ||\hat{\Delta}_{\overline{\mathbf{M}}_r}||_* - ||\hat{\Delta}_{\overline{\mathbf{M}}_r^\perp}||_* \end{aligned}$$

By the convexity of the loss function ℓ_η, together with the assumption on ρ and the definition of the dual norm,

$$\begin{aligned}\mathcal{L}_\eta(\hat{M}) - \mathcal{L}_\eta(M^*) &\geq \langle \nabla\mathcal{L}_\eta(M^*), \hat{\Delta}\rangle \geq -\left|\langle \nabla\mathcal{L}_\eta(M^*), \hat{\Delta}\rangle\right| \\ &\geq -\sigma_1(\nabla\mathcal{L}_\eta(M^*))\|\hat{\Delta}\|_* \geq \frac{\gamma}{2}\left(\|\hat{\Delta}_{\bar{\mathbf{M}}_r}\|_* + \|\hat{\Delta}_{\bar{\mathbf{M}}_r^\perp}\|_*\right).\end{aligned}$$

Since $\hat{M}$ is the optimizer of problem (7),

$$\begin{aligned}0 &\geq \left[\mathcal{L}_\eta(\hat{M}) + \gamma\|\hat{M}\|_*\right] - \left[\mathcal{L}_\eta(M^*) + \gamma\|M^*\|_*\right] \\ &\geq \left[\mathcal{L}_\eta(\hat{M}) - \mathcal{L}_\eta(M^*)\right] - \gamma\left[\|M^*\|_* - \|\hat{M}\|_*\right] \\ &\geq \frac{\gamma}{2}\|\hat{\Delta}_{\bar{\mathbf{M}}_r^\perp}\|_* - 3\|\hat{\Delta}_{\bar{\mathbf{M}}_r}\|_* - 4\|M^*_{\bar{\mathbf{M}}_r^\perp}\|_*.\end{aligned}$$

Notice that $\|M^*_{\bar{\mathbf{M}}_r^\perp}\|_* = \sum_{k=r+1}^{q}\sigma_k$, therefore the lemma holds. □

For simplicity, let

$$\frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}} = \left.\frac{\partial I_\eta(X_{ij} - M_{ij})}{\partial M_{ij}}\right|_{M_{ij}=M^*_{ij}}. \tag{12}$$

Before we look in the RSC condition, we first bound the term $\sigma_1(\nabla\mathcal{L}_\eta(M^*))$.

Lemma 7 (*Upper bound for* $\sigma_1(\nabla\mathcal{L}_\eta(M^*))$) *Suppose the noises are i.i.d with zero-mean and are symmetrically distributed around zero, then for any* $x > 0$, *and a positive constant* c_0,

$$\sigma_1(\nabla\mathcal{L}_\eta(M^*)) \leq \eta\left\{4c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x}{npq}} + \frac{8x}{3n}\right\},$$

with probability at least $1 - e^{-x}$.

Proof of Lemma 7 Since $\sigma_1(\cdot)$ is a norm, the triangle inequality holds

$$\sigma_1(\nabla\mathcal{L}_\eta(M^*)) \leq \sigma_1(\mathbb{E}[\nabla\mathcal{L}_\eta(M^*)]) + \sigma_1(\nabla\mathcal{L}_\eta(M^*) - \mathbb{E}[\nabla\mathcal{L}_\eta(M^*)]). \tag{13}$$

It can be derived from Equation (1) that

$$\nabla\mathcal{L}_\eta(M^*) = \frac{1}{2n}\sum_{t=1}^{n} J_t \sum_{i=1}^{p}\sum_{j=1}^{q} J_{t,ij}\frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}.$$

Since the noises are symmetrically distributed around zero and $\frac{\partial I_\eta(X_{ij}-M^*_{ij})}{\partial M^*_{ij}}$ is an odd function of the noise ξ_{ij}, we have $\mathbb{E}\left[\frac{\partial I_\eta(X_{ij}-M^*_{ij})}{\partial M^*_{ij}}\right] = 0$, and thus

$$\sigma_1(\mathbb{E}[\nabla\mathcal{L}_\eta(M^*)]) = 0. \tag{14}$$

To bound $\sigma_1(\nabla\mathcal{L}_\eta(M^*) - \mathbb{E}[\nabla\mathcal{L}_\eta(M^*)])$, first notice that

$$\begin{aligned}\sigma_1(\nabla L_\eta(M^*) - \mathbb{E}[\nabla L_\eta(M^*)]) &= \sup_{\substack{\|W\|_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \langle W, \nabla L_\eta(M^*) - \mathbb{E}[\nabla L_\eta(M^*)]\rangle \\ &= \sup_{\substack{\|W\|_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \frac{1}{2n}\sum_{t=1}^{\cdot}\Big\langle W, J_t\sum_{i=1}^{\cdot}\sum_{j=1}^{\cdot} J_{t,ij}\frac{\partial l_\eta(X_{ij}-M^*_{ij})}{\partial M^*_{ij}} \\ &\quad - \mathbb{E}[J_t\sum_{i=1}^{\cdot}\sum_{j=1}^{\cdot} J_{t,ij}\frac{\partial l_\eta(X_{ij}-M^*_{ij})}{\partial M^*_{ij}}]\Big\rangle \\ &:= \sup_{\substack{\|W\|_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \frac{1}{2n}\sum_{t=1}^{\cdot} f_t(M^*) \\ &:= Z.\end{aligned}$$

Since the errors $X_{ij} - M^*_{ij}$ are *i.i.d.*,

$$\sup_{\substack{\|W\|_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} f_t(M^*) \le 2\eta \quad \text{and} \quad \sup_{\substack{\|W\|_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \mathbb{E}[f_t^2(M^*)] \le \frac{\eta^2}{pq},$$

by Theorem 2.3 in Bousquet (2002), we have for any $x > 0$

$$Z \le \mathbb{E}[Z] + \sqrt{\frac{2x\eta^2}{npq} + \frac{8x\eta}{n}\mathbb{E}[Z]} + \frac{2x\eta}{3n},$$

with probability at least $1 - e^{-x}$.

Moreover, since

$$\sqrt{\frac{2x\eta^2}{npq} + \frac{8x\eta}{n}\mathbb{E}[Z]} \le \sqrt{\frac{2x\eta^2}{npq}} + \sqrt{\frac{8x\eta}{n}\mathbb{E}[Z]} \le \sqrt{\frac{2x\eta^2}{npq}} + \frac{\frac{4x\eta}{n} + 2\mathbb{E}[Z]}{2},$$

we have with probability at least $1 - e^{-x}$

$$Z \le 2\mathbb{E}[Z] + \sqrt{\frac{2x\eta^2}{npq}} + \frac{8x\eta}{3n}. \tag{15}$$

By symmetrization inequality in Boucheron et al. (2013),

$$\begin{aligned}\mathbb{E}[Z] &\le \mathbb{E}\Big[\sup_{\substack{||W||_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \frac{1}{n}\Big|\Big\langle W, \sum_{t=1}^{n} \epsilon_t J_t \sum_{i=1}^{p}\sum_{j=1}^{q} J_{t,ij} \frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}\Big\rangle\Big|\Big] \\ &= \frac{1}{n}\mathbb{E}\Big[\sup_{\substack{||W||_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \Big|\sum_{t=1}^{n} \epsilon_t \Big\langle W, J_t \sum_{i=1}^{p}\sum_{j=1}^{q} J_{t,ij} \frac{\partial I_\eta(X_{ij} - M^*_{ij})}{\partial M^*_{ij}}\Big\rangle\Big|\Big] \\ &= \frac{1}{n}\mathbb{E}\Big[\sup_{\substack{||W||_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \Big|\sum_{t=1}^{n} \epsilon_t W_{i_t j_t} \frac{\partial I_\eta(X_{i_t j_t} - M^*_{i_t j_t})}{\partial M^*_{i_t j_t}}\Big|\Big]\end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are *i.i.d.* Rademacher variables with distribution $\mathbb{P}(\epsilon_t = 1) = \mathbb{P}(\epsilon_t = -1) = \frac{1}{2}$, and are independent of $\{i_t, j_t\}_{t=1}^n$ and $\{J_t\}_{t=1}^n$.

Now, let $\mathbb{E}^*$ denote the conditional expectation given $\{X_{i_t j_t}, J_t\}_{t=1}^n$. Notice that $W_{i_t j_t} \frac{\partial I_\eta(X_{i_t j_t} - M^*_{i_t j_t})}{\partial M^*_{i_t j_t}}$ is a η–Lipschitz function of $W_{i_t j_t}$. By Theorem 4.12 in Ledoux and Talagrand (2013), we have

$$\mathbb{E}^*[Z] \le \frac{2\eta}{n}\mathbb{E}^*\Big[\sup_{\substack{||W||_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \Big|\sum_{t=1}^{n} \epsilon_t W_{i_t j_t}\Big|\Big].$$

Then take expectation over ϵ_t, and we have for a positive constant c_0,

$$\begin{aligned}\mathbb{E}[Z] &\le \frac{2\eta}{n}\mathbb{E}\Big[\sup_{\substack{||W||_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \Big|\sum_{t=1}^{n} \epsilon_t \langle J_t, W\rangle\Big|\Big] \\ &\le \frac{2\eta}{n}\mathbb{E}\Big[\sup_{\substack{||W||_* \le 1 \\ W \in \mathbb{R}^{p\times q}}} \sigma_1\Big(\sum_{t=1}^{n} \epsilon_t J_t\Big)\|W\|_*\Big] \\ &\le 2\eta\mathbb{E}[\sigma_1(\frac{1}{n}\sum_{t=1}^{n} \epsilon_t J_t)] \\ &\le 2\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\log(q+1)}\frac{\log(p+q)}{n}\right],\end{aligned} \tag{16}$$

where the second inequality follows from the definition of dual norm, and the last inequality follows from Proposition 2 in Koltchinskii et al. (2011): it is simple to show that

$$\max\{\sigma_1(\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[J_t J_t^\top]), \sigma_1(\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[J_t^\top J_t])\} = \frac{1}{q},$$

besides, since $\sigma_1(\epsilon_t J_t) = |\epsilon_t|\sigma_1(J_t) \le |\epsilon_t|$, we have

$$U^{(2)}_{\epsilon_t J_t} \le U^{(2)}_{\epsilon_t} = \sqrt{\frac{1}{\log 2}},$$

where $U^{(\alpha)}_Z$ is defined as $U^{(\alpha)}_Z = \inf\{u > 0 : \mathbb{E}\exp(\frac{\sigma_1(Z)^\alpha}{u^\alpha}) \le 2\}$, then by concavity of logarithm, we have

$$\sqrt{\log\frac{q}{\log 2}} = \sqrt{\frac{1}{2}\log q + \frac{1}{2}\log(\frac{1}{\log 2})} \le \sqrt{\log(\frac{q}{2} + \frac{1}{2\log 2})} \le \sqrt{\log(q+1)},$$

finally, using Proposition 2 in Koltchinskii et al. (2011), we have $\forall \tilde{x} > 0$ and a constant $\tilde{c}_0$

$$\mathbb{P}\left\{\sigma_1(\frac{1}{n}\sum_{t=1}\epsilon_t J_t) \ge c_0\left[\sqrt{\frac{\tilde{x}+\log(p+q)}{nq}} + \sqrt{\log(q+1)}\frac{\tilde{x}+\log(p+q)}{n}\right]\right\} \le e^{-\tilde{x}}.$$

Then

$$\begin{aligned}\mathbb{E}[\sigma_1(\frac{1}{n}\sum_{t=1}\epsilon_t J_t)] &= \int_0^\infty \mathbb{P}(\sigma_1(\frac{1}{n}\sum_{t=1}\epsilon_t J_t) \ge s)ds \\ &\le c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\log(q+1)}\frac{\log(p+q)}{n}\right] \\ &\quad + c_0\int_0^\infty e^{-\tilde{x}}\left(\frac{1}{2\sqrt{nq(\tilde{x}+\log(p+q))}} + \frac{\sqrt{\log(q+1)}}{n}\right)d\tilde{x},\end{aligned}$$

since $\frac{1}{\sqrt{\tilde{x}+\log(p+q)}} \le \frac{\sqrt{2}}{2}\left(\frac{1}{\sqrt{\tilde{x}}} + \frac{1}{\sqrt{\log(p+q)}}\right)$, after simplification, we have

$$\mathbb{E}[\sigma_1(\frac{1}{n}\sum_{t=1}\epsilon_t J_t)] \le c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\log(q+1)}\frac{\log(p+q)}{n}\right], \tag{17}$$

where c_0 is a constant independent of n, p and q.

By Equation (13), together with Equation (14), (15), (16) and (17), we have with probability at least $1 - e^{-x}$

$$\sigma_1(\nabla \mathcal{L}_\eta(M^*)) \le \eta\left\{4c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\log(q+1)}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x}{npq}} + \frac{8x}{3n}\right\}.$$

□

Proof of Lemma 2 $\delta\mathcal{L}_\eta(M, M^*)$ can be written as

$$\begin{aligned}\delta L_\eta &= \delta L_\eta + \mathbb{E}[\delta L_\eta] - \mathbb{E}[\delta L_\eta] \\ &\geq \mathbb{E}[\delta L_\eta] - |\mathbb{E}[\delta L_\eta] - \delta L_\eta|.\end{aligned}$$

In the following, we establish the lower bound for $\mathbb{E}[\delta L_\eta(M, M^*)]$ and the upper bound for $|\mathbb{E}[\delta L_\eta(M, M^*)] - \delta L_\eta(M, M^*)|$, respectively, for $M \in \{M : \|M - M^*\|_{\max} \leq \eta\} \cap \mathbb{M}_0$.

Given any $M \in \{M : \|M - M^*\|_{\max} \leq \eta\} \cap \mathbb{M}_0$ and $\Delta = M - M^*$,

$$\begin{aligned}\mathbb{E}[\delta L_\eta(M, M^*)] &= \frac{1}{2n}\sum_{t=1}^{n}\mathbb{E}\{\sum_{i=1}^{p}\sum_{j=1}^{q} J_{t,ij}[I_\eta(X_{ij} - M_{ij}) - I_\eta(X_{ij} - M_{ij}^*) - \frac{\partial I_\eta(X_{ij} - M_{ij}^*)}{\partial M_{ij}^*}\Delta_{ij}]\} \\ &= \frac{1}{2pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\mathbb{E}[I_\eta(X_{ij} - M_{ij})] - \mathbb{E}[I_\eta(X_{ij} - M_{ij}^*)] - \mathbb{E}[\frac{\partial I_\eta(X_{ij} - M_{ij}^*)}{\partial M_{ij}^*}]\Delta_{ij},\end{aligned}$$

where $\frac{\partial I_\eta(X_{ij} - M_{ij}^*)}{\partial M_{ij}^*}$ is defined in Equation (12).

Since $I_\eta(X_{ij} - M_{ij})$ and $\frac{\partial I_\eta(X_{ij} - M_{ij})}{\partial M_{ij}^*}$ are continuous function of M_{ij},

$$\begin{aligned}\mathbb{E}[\frac{\partial I_\eta(X_{ij} - M_{ij})}{\partial M_{ij}}] &= \frac{\partial \mathbb{E}[I_\eta(X_{ij} - M_{ij})]}{\partial M_{ij}} \\ &= \int_{|M_{ij} - X_{ij}| \leq \eta}(M_{ij} - X_{ij})dF(X_{ij}) + \eta\int_{M_{ij} - X_{ij} > \eta} dF(X_{ij}) - \eta\int_{M_{ij} - X_{ij} < -\eta} dF(X_{ij}) \\ &= (M_{ij} - X_{ij})F(X_{ij})\Big|_{M_{ij} - \eta}^{M_{ij} + \eta} - \int_{M_{ij} - \eta}^{M_{ij} + \eta} F(X_{ij})d(-X_{ij}) \\ &\quad + \eta F(M_{ij} - \eta) - \eta[1 - F(M_{ij} + \eta)] \\ &= \int_{M_{ij} - \eta}^{M_{ij} + \eta} F(X_{ij})dX_{ij} - \eta,\end{aligned} \tag{18}$$

where $F(\cdot)$ is the *cdf* of X_{ij}, and

$$\frac{\partial^2 \mathbb{E}[I_\eta(X_{ij} - M_{ij})]}{\partial M_{ij}^2} = F(M_{ij} + \eta) - F(M_{ij} - \eta)$$

Apply Taylor's theorem to $\mathbb{E}[I_\eta(X_{ij} - M_{ij})]$, and we have for some $t_{ij} \in (0, 1)$

$$\begin{aligned}\mathbb{E}[\delta L_\eta(M, M^*)] &= \frac{1}{2pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{1}{2}[F(M_{ij}^* + t_{ij}\Delta_{ij} + \eta) - F(M_{ij}^* + t_{ij}\Delta_{ij} - \eta)]\Delta_{ij}^2 \\ &= \frac{1}{4pq}\sum_{i=1}^{p}\sum_{j=1}^{q}[F_\xi(t_{ij}\Delta_{ij} + \eta) - F_\xi(t_{ij}\Delta_{ij} - \eta)]\Delta_{ij}^2 \\ &\geq \frac{1}{4c_1^2 pq}\|\Delta\|_F^2,\end{aligned} \tag{19}$$

where the inequality follows from the Assumption 2.

Next, we consider the upper bound for $\left|\mathbb{E}[\delta L_\eta] - \delta L_\eta\right|$. The techniques used here are similar to those in the proof Lemma 7.

Let $\delta I_{\eta,ij} = I_\eta(X_{ij} - M_{ij}) - I_\eta(X_{ij} - M_{ij}^*) - \frac{\partial I_\eta(X_{ij} - M_{ij}^*)}{\partial M_{ij}^*}\Delta_{ij}$, since I_η is $\eta-$Lipschitz,

$$\|\delta l_{\eta,ij}\|_\infty \le 2\eta\|\Delta\|_\infty \le 2\eta\|\Delta\|_F. \tag{20}$$

For any $M \in \mathrm{M}_0$, let $\Delta = M - M^*$, we have

$$\begin{aligned}\left|\mathbb{E}[\delta L_\eta(M, M^*)] - \delta_\eta(\hat{M}, M^*)\right| &= \frac{1}{2n}\left|\sum_{t=1}^n (\delta l_{\eta,i_tj_t} - \mathbb{E}[\delta l_{\eta,i_tj_t}])\right| \\ &:= \frac{1}{2n}\left|\sum_{t=1}^n f_{1t}(M)\right| = \frac{\|\Delta\|_F}{2n}\left|\sum_{t=1}^n \frac{f_{1t}(M)}{\|\Delta\|_F}\right| \\ &\le \frac{\|\Delta\|_F}{2n}\sup_{M\in\mathrm{M}_0}\left|\sum_{t=1}^n \frac{f_{1t}(M)}{\|\Delta\|_F}\right|.\end{aligned} \tag{21}$$

Let $Z_1 = \frac{1}{n}\sup_{M\in\mathbb{R}^{p\times q}}\left|\sum_{t=1}^n \frac{f_{1t}(M)}{\|\Delta\|_F}\right|$. By Equation (20), $\frac{f_{1t}(M)}{\|\Delta\|_F} \le 4\eta$ and $\mathbb{E}[\frac{f_{1t}^2(M)}{\|\Delta\|_F^2}] \le \frac{\eta^2}{pq}$ for any $M \in \mathbb{R}^{p\times q}$. Since the errors $X_{ij} - M_{ij}^*$ are *i.i.d.*, by Theorem 2.3 in Bousquet (2002), for any $x \ge 0$

$$Z_1 \le 2\mathbb{E}(Z_1) + 2\sqrt{\frac{2x\eta^2}{npq} + \frac{16x\eta}{3n}}, \tag{22}$$

with probability at least $1 - e^{-x}$.

Let ϵ_t's be *i.i.d.* Rademacher variables. Then, by symmetrization inequality in Boucheron et al. (2013),

$$\begin{aligned}\mathbb{E}[Z_1] &= \frac{1}{n}\mathbb{E}\sup_{M\in\mathrm{M}_0}\left|\sum_{t=1}^n \frac{f_{1t}(M)}{\|\Delta\|_F}\right| \\ &\le \frac{2}{n}\mathbb{E}\sup_{M\in\mathrm{M}_0}\left|\sum_{t=1}^n \epsilon_t \frac{\delta l_{\eta,i_tj_t}}{\|\Delta\|_F}\right|\end{aligned}$$

Let $\mathbb{E}^*$ denote the conditional expectation given $\{i_t, j_t\}_{t=1}^n$.

By contraction principle in Theorem 4.4 of Ledoux and Talagrand (2013), since $|\frac{\delta l_{\eta,i_tj_t}}{\Delta_{i_tj_t}}| \le 2\eta$,

$$\mathbb{E}^*\left[\sup_{M\in\mathrm{M}_0}\left|\sum_{t=1}^n \epsilon_t \frac{\delta l_{\eta,i_tj_t}}{\Delta_{i_tj_t}}\frac{\Delta_{i_tj_t}}{\|\Delta\|_F}\right|\right] \le 4\eta\mathbb{E}^*\left[\sup_{M\in\mathrm{M}_0}\left|\sum_{t=1}^n \epsilon_t \frac{\Delta_{i_tj_t}}{\|\Delta\|_F}\right|\right].$$

Then

$$\begin{aligned}\mathbb{E}[Z_1] &\le \frac{8\eta}{n}\mathbb{E}\sup_{M\in\mathrm{M}_0}\left|\sum_{t=1}^n \epsilon_t \frac{\langle J_t, \Delta\rangle}{\|\Delta\|_F}\right| \\ &\le 8\eta\mathbb{E}\sup_{M\in\mathrm{M}_0}\frac{\|\sum_{t=1}^n \frac{1}{n}\epsilon_t J_t\|_1\|\Delta\|_*}{\|\Delta\|_F} \\ &\le 8\eta\mathbb{E}\ \sigma_1\left(\sum_{t=1}^n \frac{1}{n}\epsilon_t J_t\right)\sup_{M\in\mathrm{M}_0}\frac{\|\Delta\|_*}{\|\Delta\|_F}.\end{aligned}$$

The second inequality follows form the definition of the dual norm.

By Lemma 1, we have for $M \in \mathrm{M}_0$ and $r > 0$,

$$||\Delta_{\overline{\mathrm{M}}_r^{\perp}}||_* \le 3||\Delta_{\overline{\mathrm{M}}_r}||_* + 4\sum_{k=r+1}^{q}\sigma_k.$$

Note that

$$\begin{aligned}\operatorname{rank}(\Delta_{\overline{\mathrm{M}}_r}) &= \operatorname{rank}(\Delta - \Delta_{\overline{\mathrm{M}}_r^{\perp}}) \\ &= \operatorname{rank}(U_r U_r^{\top}\Delta(I - V_r V_r^{\top}) + \Delta V_r V_r^{\top}) \le 2r.\end{aligned}$$

Then we have for any $M \in \mathbb{R}^{p\times q}$ and $\Delta = M - M^*$

$$\begin{aligned}\|\Delta\|_* &\le \|\Delta_{\overline{\mathrm{M}}_r}\|_* + \|\Delta_{\overline{\mathrm{M}}_r^{\perp}}\|_* \le 4\|\Delta_{\overline{\mathrm{M}}_r}\|_* + 4\sum_{k=r+1}^{q}\sigma_k \\ &\le 4\sqrt{2r}\|\Delta\|_F + 4\sum_{k=r+1}^{q}\sigma_k.\end{aligned} \tag{23}$$

If M^* is exactly low-rank with $\operatorname{rank}(M^*) \le r$, then $\sum_{k=r+1}^{q}\sigma_k = 0$, in this case

$$\begin{aligned}\mathbb{E}[Z_1] &\le 32\sqrt{2r}\eta\mathbb{E}\left[\sigma_1\left(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t J_t\right)\right] \\ &\le 32\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right],\end{aligned}$$

where the last inequality follows from Equation (17).

Then, by Equation (22)

$$Z_1 \le 64\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + 2\sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x\eta^2}{npq}} + \frac{16x\eta}{3n},$$

with probability at least $1 - e^{-x}$.

Therefore, by Equation (21), with probability at least $1 - e^{-x}$.

$$\begin{aligned}|\mathcal{L}_\eta - \mathbb{E}[\mathcal{L}_\eta]| &\le Z_1\|\Delta\|_F/2 \\ &\le \left\{32\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x\eta^2}{npq}} + \frac{8x\eta}{3n}\right\}\|\hat{\Delta}\|_F.\end{aligned}$$

Together with Equation (19), we have with probability at least $1 - e^{-x}$,

$$\delta \mathrm{L}_\eta \geq \frac{1}{4c_1^2 pq}\|\Delta\|_F^2 - \left\{32\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x\eta^2}{npq}} + \frac{8x\eta}{3n}\right\}\|\Delta\|_F \tag{24}$$

□

Proof of Theorem 1 Construct $M_t = M^* + t(\hat{M} - M^*)$ in the following way. If $\|\hat{M} - M^*\|_{\max} < \eta$, then $t = 1$, otherwise, choose t such that $\|M_t - M^*\|_{\max} = \eta$. Let $\Delta_t = M_t - M^* = t(\hat{M} - M^*) = t\hat{\Delta}$. Notice

$$\begin{aligned}\delta \mathrm{L}_\eta(M_t) &= \mathrm{L}_\eta(M_t) - \mathrm{L}(M^*) - \langle \nabla \mathrm{L}(M^*), \Delta_t\rangle \\ &\leq t\mathrm{L}(\hat{M}) + (1-t)\mathrm{L}(M^*) - \mathrm{L}(M^*) - \langle \nabla \mathrm{L}(M^*), \Delta_t\rangle \\ &= t\delta \mathrm{L}_\eta(\hat{M}).\end{aligned}$$

Since $\hat{M}$ is the optimizer of problem (7),

$$\mathrm{L}_\eta(\hat{M}) + \gamma\|\hat{M}\|_* - \left[\mathrm{L}_\eta(M^*) + \gamma\|M^*\|_*\right] \leq 0$$

Therefore,

$$\begin{aligned}\delta \mathrm{L}_\eta &= \mathrm{L}_\eta(\hat{M}) - \mathrm{L}_\eta(M^*) - \langle \nabla \mathrm{L}_\eta(M^*), \Delta\rangle \\ &\leq \gamma(\|M^*\|_* - \|\hat{M}\|_*) + \langle \nabla \mathrm{L}_\eta(M^*), \Delta\rangle \\ &\leq \gamma\|\Delta\|_* - \sigma_1(\nabla \mathrm{L}_\eta(M^*))\|\Delta\|_* \\ &\leq \frac{3}{2}\gamma\|\Delta\|_*\end{aligned}$$

Then by Lemma 2, for any $x > 0$, with probability at least $1 - e^{-x}$

$$\frac{1}{4c_1^2 pq}\|\Delta\|_F^2 \leq \frac{3}{2}\gamma\|\Delta\|_* + \left\{32\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}} + \sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right] + \sqrt{\frac{2x\eta^2}{npq}} + \frac{8x\eta}{3n}\right\}\|\Delta\|_F$$

Divided both sides of the inequality by $\|\Delta\|_F$, we have

$$\frac{1}{4c_1^2pq}\|\nabla\mathcal{L}_\eta(M^*)\|_2 \le \frac{3}{2}\gamma\|\nabla\mathcal{L}_\eta(M^*)\|_2 + \left\{32\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}}+\sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right]+\sqrt{\frac{2x\eta^2}{npq}}+\frac{8x\eta}{3n}\right\}$$
$$\le 6\gamma\sqrt{2r} + \left\{32\sqrt{2r}\eta c_0\left[\sqrt{\frac{\log(p+q)}{nq}}+\sqrt{\frac{\log(q+1)}{n}}\frac{\log(p+q)}{n}\right]+\sqrt{\frac{2x\eta^2}{npq}}+\frac{8x\eta}{3n}\right\},$$

with probability at least $1-e^{-x}$. The second inequality follows from Equation (23) when M^* has rank smaller than r and the fact that $2\sigma_1(\nabla\mathcal{L}(M^*))$ with probability at least $1-e^{-x}$ by Lemma 7.

Take $x=\log(p+q)$ and $n > C(L)\ c_1^2\sqrt{2rp\log(p+q)\log(q+1)}$ with $C(L)$ with being some constant depending on L, we have $\|\Delta_t\|_{\max} \le \frac{L}{\sqrt{pq}}\|\Delta_t\|_F < \eta$. Then by the construction of M_t, $t=1$. Finally, we have with probability at least $2e^{-x}-e^{-2x}=1-3(p+q)^{-1}$,

$$\frac{1}{\sqrt{pq}}\|\hat{\Delta}\|_F \le C_1\cdot c_1^2\eta\sqrt{\frac{p\log(p+q)\log(q+1)}{n}}\left(\sqrt{2r}c_2+c_3\right),$$

where C_1, c_2, c_3 are absolute constants. □

Appendix 3: Proof for reduced rank regression

Lemma 8 (*Upper Bound for* $\sigma_1(\nabla\mathcal{L}_\eta(C^*))$) *Suppose that* ξ_{ij}*'s are i.i.d. with zero mean and symmetrically distributed around zero, then for any* $x>0$, *we have with probability at least* $1-e^{-x}$,

$$\sigma_1(\nabla\mathcal{L}_\eta(C^*)) \le 4\eta\sigma_1(\Sigma)\left(\sqrt{4n(p+q)+2nx}+2(p+q)+x\right)$$

Proof of Lemma 8 Note $\frac{\partial}{\partial c_{kj}}\mathcal{L}(C) = \sum_{i=1}^n -\ell_\eta'(y_{ij}-x_i^\top c_j)x_{ik}$. Let $g_{ij}=\ell_\eta'(y_{ij}-x_i^\top c_j)$, $G=[g_{ij}]_{n\times q}$ and G^* the value of G when $C=C^*$, then $\nabla\mathcal{L}_\eta(C^*) = X^\top G^*$.

$$\sigma_1(\nabla\mathcal{L}_\eta(C^*)) = \sup_{\substack{\|u\|_2=1\\ u\in\mathbb{R}^q}}\ \sup_{\substack{\|v\|_2=1\\ v\in\mathbb{R}^p}}\langle Xv, G^*u\rangle.$$

Following the proof of Lemma 3 Negahban and Wainwright (2011) (the proof is given in its supplementary material), we have

$$\mathbb{P}\left(\|\nabla\mathcal{L}(X^\top G^*)\|_2 \ge 4\delta n\right) \le 8^{p+q}\max_{\substack{\|u\|_2=1\\ u\in\mathbb{R}^q}}\max_{\substack{\|v\|_2=1\\ v\in\mathbb{R}^p}}\mathbb{P}\left(\frac{\langle Xv, G^*u\rangle}{n}\ge\delta\right). \tag{25}$$

It remains to bound $\frac{1}{n}\langle Xv, G^*u\rangle$. Let

$$Z := \frac{1}{n}\langle Xv, G^*u\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle v, x_i\rangle\langle u, g_i^*\rangle,$$

where g_i^* is the i-th row of G^*. Since ξ_{ij}'s are symmetrically distributed around zero and $l_\eta'(x)$ is an odd function, $\mathbb{E}[G^*] = 0$. Hence, $\mathbb{E}\{\langle v, x_i\rangle\langle u, g_i^*\rangle\} = 0$. Further, for k being any positive integer,

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\{\langle v, x_i\rangle^{2k}\langle u, g_i^*\rangle^{2k}\} &\le \eta^{2k}\mathbb{E}\sum_{i=1}^{n}\langle v, x_i\rangle^{2k}\\ &= \eta^{2k}\mathbb{E}\sum_{i=1}^{n}(x_i^\top u)^{2k}\\ &= \eta^{2k}n(u^\top\Sigma u)^{2k}(2k-1)!!\\ &\le \eta^{2k}n(2k-1)!!\sigma_1(\Sigma)^{2k}.\end{aligned}$$

By Berstein's inequality, for any $t > 0$ and u, v satisfying $\|u\|_2 = 1, \|v\|_2 = 1$,

$$\mathbb{P}\{Z \ge \eta\sigma_1(\Sigma)(\sqrt{2t/n} + t/n)\} \le e^{-t}.$$

Combining with Equation (25), we have

$$\mathbb{P}\left(\|\nabla\mathcal{L}_\eta(X^\top G^*)\| \ge 4n\eta\sigma_1(\Sigma)(\sqrt{2t/n} + t/n)\right) \le 8^{p+q}e^{-t}.$$

Take $t = 2(p+q) + x$ for any $x > 0$, and then we have

$$\mathbb{P}\left(\|\nabla\mathcal{L}_\eta(X^\top G^*)\| \ge 4\eta\sigma_1(\Sigma)(\sqrt{4n(p+q) + 2nx} + 2(p+q) + x)\right) \le e^{-x}.$$

□

$$\begin{aligned}\delta\mathcal{L}_\eta &= \delta\mathcal{L}_\eta + \mathbb{E}[\delta\mathcal{L}_\eta] - \mathbb{E}[\delta\mathcal{L}_\eta]\\ &\ge \mathbb{E}[\delta\mathcal{L}_\eta] - \left|\delta\mathcal{L}_\eta - \mathbb{E}[\delta\mathcal{L}_\eta]\right|.\end{aligned}$$

Proof of Lemma 5

In the following, we establish the lower bound for $\mathbb{E}[\delta\mathcal{L}_\eta]$ and the upper bound for $\left|\delta\mathcal{L}_\eta - \mathbb{E}[\delta\mathcal{L}_\eta]\right|$, respectively, for $C \in \mathcal{C}_0 \cap \{C : \|C - C^*\|_F \le \eta\}$

Given any $C \in \mathcal{C}_0 \cap \{C : \|C - C^*\|_F \le \eta\}$ and $\Delta = C - C^*$, for some $t_{ij} \in (0, 1)$

$$\begin{aligned}\mathbb{E}[\delta\mathcal{L}_\eta] &= \sum_{i=1}^{n}\sum_{j=1}^{m}\{\mathbb{E}[l_\eta(y_{ij} - x_i^\top c_j)] - \mathbb{E}[l_\eta(y_{ij} - x_i^\top c_j^*)] - \mathbb{E}[l_\eta'(y_{ij} - x_i^\top c_j^*)(-x_i^\top\Delta_j)]\}\\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}_X[\{F_\xi(t_{ij}x_i^\top\Delta_j + \eta) - F_\xi(t_{ij}x_i^\top\Delta_j - \eta)\}(x_i^\top\Delta_j)^2]\\ &\ge \frac{1}{2c_1^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}(x_i^\top\Delta_j)^2\\ &\ge \frac{n}{2c_1^2}\sigma_n(\Sigma)\sum_{j=1}^{m}\|\Delta_j\|_F^2 = \frac{n\sigma_n(\Sigma)}{2c_1^2}\|\Delta\|_F^2,\end{aligned}$$

where the equality follows from Taylor's theorem, and the first inequality follows from Assumption 2 and Assumption 4. For the calculation of $\frac{\partial^2 \mathbb{E}[l_\eta(y_{ij}-\langle Z^{ij},C\rangle)]}{\partial(\langle Z^{ij},C\rangle)^2}$, please refer to the calculation of $\frac{\partial^2 \mathbb{E}[l_\eta(X_{ij}-M_{ij}]}{\partial M_{ij}^2}$ in the case of matrix completion problems.

For any $i=1,\dots,n, j=1,\dots,q$, there exist $\tau_{ij}\in(0,1)$, such that $\ell_\eta(y_{ij}-x_i^\top c_j)-\ell_\eta(y_{ij}-x_i^\top c_j^*)=\ell'_\eta(y_{ij}-x_i^\top \tilde{c}_j)x_i^\top(c_j^*-c_j)$, where $\tilde{c}_j=c_j^*+\tau_{ij}(c_j-c_j^*)$. Therefore,

$$\delta L_\eta(C)=\langle \nabla L_\eta(\tilde{C})-\nabla L_\eta(C^*), C-C^*\rangle.$$

Then

$$\begin{aligned}\big|\delta L_\eta-\mathbb{E}[\delta L_\eta]\big| &= |\langle \nabla L_\eta(\tilde{C})-\nabla L_\eta(C^*), C-C^*\rangle-\mathbb{E}\{\langle \nabla L_\eta(\tilde{C})-\nabla L_\eta(C^*), C-C^*\rangle\}|\\ &= |\langle X^\top\tilde{G}-X^\top G^*, C-C^*\rangle-\mathbb{E}\{\langle X^\top\tilde{G}-X^\top G^*, C-C^*\rangle\}|\\ &\le \|\Delta\|_*\sigma_1(X^\top(\tilde{G}-G^*)-\mathbb{E}(X^\top(\tilde{G}-G^*))).\end{aligned}$$

Following the proof in Lemma 8, we have for any $x>0$

$$\sigma_1(X^\top(\tilde{G}-\mathbb{E}\tilde{G})-G^*))\le 12\eta\sigma_1(\Sigma)\Big(\sqrt{4n(p+q)+2nx}+2(p+q)+x\Big),$$

with probability at least $1-e^{-x}$.

Similar to Equation (23), it can be shown that if C^* has rank smaller than r, then $\sup_{C\in C_0}\frac{\|\Delta\|_*}{\|\Delta\|_F}\le 4\sqrt{2r}$. Hence, for $C\in C_0, \|\Delta\|_*\le 4\sqrt{2r}\|\Delta\|_F$. Now we have with probability at least $1-e^{-x}$,

$$\delta L_\eta\ge\frac{n\sigma_n(\Sigma)}{2c_1^2}\|\Delta\|_F^2-48\sqrt{2r}\eta\sigma_1(\Sigma)\Big(\sqrt{4n(p+q)+2nx}+2(p+q)+x\Big)\|\Delta\|_F$$

□

Proof of Theorem 2 Construct $C_t=C^*+t(\hat{C}-C^*)$ in the following way. If $\|\hat{C}-C^*\|_F<\eta$, then $t=1$, otherwise, choose t such that $\|C_t-C^*\|_F=\eta$. Let $\Delta_t=C_t-C^*=t(\hat{C}-C^*)=t\hat{\Delta}$. Notice

$$\begin{aligned}\delta L_\eta(C_t) &= L_\eta(C_t)-L_\eta(C^*)-\langle \nabla L_\eta(C^*),\Delta_t\rangle\\ &\le tL_\eta(\hat{C})+(1-t)L_\eta(C^*)-L_\eta(C^*)-\langle \nabla L_\eta(C^*),\Delta_t\rangle\\ &= t\delta L_\eta(\hat{C}).\end{aligned}$$

Since $\hat{C}$ is the optimizer of problem (9), we have

$$L_\eta(\hat{C})+\gamma\|\hat{C}\|_*\le L_\eta(C^*)+\gamma\|C^*\|_*.$$

Therefore,

$$\begin{aligned}\delta L_\eta(\hat{C}) &= L_\eta(\hat{C}) - L_\eta(C^*) - \langle \nabla L_\eta(C^*), \hat{\Delta} \rangle \\ &\le \gamma(\|C^*\|_* - \|\hat{C}\|_*) + \langle \nabla L_\eta(C^*), \hat{\Delta} \rangle \\ &\le \gamma\|\hat{\Delta}\|_* + \sigma_1(\nabla L_\eta(C^*))\|\hat{\Delta}\|_* \\ &\le \frac{3}{2}\gamma\|\hat{\Delta}\|_*.\end{aligned}$$

By Lemma 5, for any $x > 0$, with probability at least $1 - e^{-x}$,

$$\frac{3t}{2}\gamma\|\hat{\Delta}\|_* + 48\sqrt{2r}\eta\sigma_1(\Sigma)\left(\sqrt{4n(p+q) + 2nx} + 2(p+q) + x\right)\|\Delta_t\|_F \ge \frac{n\sigma_n(\Sigma)}{2c_1^2}\|\Delta_t\|_F^2.$$

The second inequality follows from Equation (23) when M^* has rank smaller than r and the fact that the selection of $\gamma \ge 2\sigma_1(\nabla L(M^*))$ with probability at least $1 - e^{-x}$ by Lemma 7.

Further, by Equation (23) and the fact that $t \ge 2\sigma_1(\nabla L(C^*))$ with probability at least $1 - e^{-x}$ by Lemma 8, we have

$$\frac{n\sigma_n(\Sigma)}{2c_1^2}\|\Delta_t\|_F \le \frac{3}{2}\gamma 4\sqrt{2r} + 48\sqrt{2r}\eta\sigma_1(\Sigma)\left(\sqrt{4n(p+q) + 2nx} + 2(p+q) + x\right)$$

with probability at least $(1 - e^{-x})^2$. Take $x = p + q$ and $n > C_2 \cdot \frac{\sigma_1(\Sigma)}{\sigma_n(\Sigma)} c_1^2(p+q)r$ with C_2 being some constant, we have $\|\Delta_t\|_F < \eta$. Then by the construction of t, $t = 1$. Finally, we have with probability at least $1 - 2e^{-x} - e^{-2x} = 1 - 3e^{-(p+q)}$,

$$\frac{1}{2c_1^2}\|\Delta\|_F \le 48\sqrt{2r}\eta\frac{\sigma_1(\Sigma)}{\sigma_n(\Sigma)}\left(\sqrt{\frac{6(p+q)}{n}} + \frac{3(p+q)}{n}\right).$$

□

Author contributions All authors contributed to the conception and design in methods, theory, and algorithms. Theoretical development were performed by NJ, and the experimental evaluation was performed by EXF. All authors participated in preparing, reading, and revising the manuscript; all authors approved the manuscript.

Funding Tang was supported in part by a Subaward of an NIH Grant R01GM140476, and an NSF Grant DMS-2210687. Fang was partially supported by NSF Grants DMS-1820702, DMS-1953196, DMS-2015539, and a Grant from Whitehead foundation.

Data availibility The real data sets to evaluate the performance of the methods in this paper are publicly available. 'Jester Joke' data set is available through http://eigentaste.berkeley.edu/dataset/, and 'Cameraman image' data is available through http://sipi.usc.edu/database/database.cgi?volume=misc&image=8#top.

Code availability The MATLAB code is available upon request to the corresponding author.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Agarwal, A., Negahban, S., & Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 1171–1197.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. New York: Wiley, 3rd edition.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique, 334*(6), 495–500.

Candès, E. J., & Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory, 56*(5), 2053–2080.

Charisopoulos, V., Chen, Y., Davis, D., Díaz, M., Ding, L., & Drusvyatskiy, D. (2021). Low-rank matrix recovery with composite optimization: Good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 1–89.

Chen, C., He, B., & Yuan, X. (2012). Matrix completion via an alternating direction method. *IMA Journal of Numerical Analysis, 32*(1), 227–245.

Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics,* (Vol. 482). Hoboken: John Wiley & Sons.

Elsener, A., & van de Geer, S. (2018). Robust low-rank matrix estimation. *Annals of Statistics, 46*(6B) 3481–3509.

Fan, J., Wang, W., & Zhu, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of Statistics, 49*(3), 1239.

Freund, R. M., & Grigas, P. (2016). New analysis and results for the Frank–Wolfe method. *Mathematical Programming, 155*(1–2), 199–230.

Freund, R. M., Grigas, P., & Mazumder, R. (2017). An extended Frank–Wolfe method with "in-face'' directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization, 27*(1), 319–346.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: The approach based on influence functions* (Vol. 196). Hoboken: John Wiley & Sons.

Huber, P. J. (2004). *Robust statistics* (Vol. 523). Hoboken: John Wiley & Sons.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International conference on machine learning* (pp. 427–435). PMLR.

Kerdreux, T., d'Aspremont, A., & Pokutta, S. (2018). Restarting frank-wolfe. arXiv preprint arXiv: 1810. 02429.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli, 20*(1), 282–303.

Koltchinskii, V., Lounici, K., & Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics, 39*(5), 2302–2329.

Lacoste-Julien, S. & Jaggi, M. (2015). On the global linear convergence of Frank–Wolfe optimization variants. arXiv preprint arXiv: 1511. 05932.

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Oxford: Clarendon Press.

Ledoux, M., & Talagrand, M. (2013). *Probability in Banach spaces: Isoperimetry and processes*. Springer, Berlin.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics, 45*(2), 866–896.

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

Negahban, S., & Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research, 13*(1), 1665–1697.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., & Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science, 27*(4), 538–557.

Recht, B. (2011). A simpler approach to matrix completion. *Journal of Machine Learning Research*, *12*(12).

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., & Smola, A. (2016). Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning* (pp. 314–323).

Reinsel, G. C., & Velu, R. (1998). *Multivariate reduced rank regression*. Berlin: Springer.

Rohde, A., & Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics, 39*(2), 887–930.

She, Y., & Chen, K. (2017). Robust reduced-rank regression. *Biometrika, 104*(3), 633–647.

Sun, Q., Zhou, W.-X., & Fan, J. (2020). Adaptive Huber regression. *Journal of the American Statistical Association, 115*(529), 254–265.

Swoboda, P., & Kolmogorov, V. (2019). Map inference via block-coordinate Frank–Wolfe algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11146–11155).
Toh, K.-C., & Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization, 6*(615–640), 15.
Wong, R. K., & Lee, T. C. (2017). Matrix completion with noisy entries and outliers. *The Journal of Machine Learning Research, 18*(1), 5404–5428.
Zhou, W.-X., Bose, K., Fan, J., & Liu, H. (2018). A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *The Annals of Statistics, 46*(5), 1904–1931.