Bernoulli **30**(3), 2024, 2326–2345
https://doi.org/10.3150/23-BEJ1675

Low-rank matrix recovery under heavy-tailed errors

MYEONGHUN YU[1,a], QIANG SUN[2,b] and WEN-XIN ZHOU[3,c]

[1]*Department of Mathematics, University of California, San Diego, La Jolla, CA, 92093, USA,* [a]*myyu@ucsd.edu*
[2]*Department of Statistical Sciences, University of Toronto, Toronto, ON M5G 1Z5, Canada,*
[b]*qiang.sun@utoronto.ca*
[3]*Department of Information and Decision Sciences, University of Illinois at Chicago, Chicago, IL, 60607, USA,*
[c]*wenxinz@uic.edu*

This paper proposes convex relaxation based robust methods to recover approximately low-rank matrices in the presence of heavy-tailed and asymmetric errors, allowing for heteroscedasticity. We focus on three archetypal applications in matrix recovery: matrix compressed sensing, matrix completion and multitask regression. Statistically, we provide sub-Gaussian-type deviation bounds when the noise variables only have bounded variances in each aforementioned setting. Improving upon the earlier results in Fan, Wang and Zhu (*Ann. Statist.* **49** (2021) 1239–1266), the convergence rates of our estimators are proportional to the noise scale under matrix sensing and multitask regression settings, and thus diminish to 0 in the noiseless case. Computationally, we propose a matrix version of the local adaptive majorize-minimization algorithm, which is much faster than the alternating direction method of multiplier used in previous work and is scalable to large datasets. Numerical experiments demonstrate the advantage of our methods over their non-robust counterparts and corroborate the theoretical findings that the convergence rates are proportional to the noise scale.

Keywords: Heavy-tailed data; Huber loss; low-rank matrix recovery; nuclear norm; trace regression

1. Introduction

There has been a recent surge of interest in matrix recovery which aims to recover an unknown matrix from noisy observations. Matrix recovery has wide applications in practice, including collaborative filtering [14], multitask regression [1], quantum state tomography [15] and face recognition [24], to name a few. Statistically, it aims to estimate $\mathbf{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ based on n independently and identically distributed (i.i.d.) observations $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ following the generative model

$$y_i = \operatorname{tr}(\boldsymbol{X}_i^{\mathrm{T}} \mathbf{\Theta}^*) + \epsilon_i =: \langle \boldsymbol{X}_i, \mathbf{\Theta}^* \rangle + \epsilon_i,$$

where $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ is a random measurement matrix, and ϵ_i is an error variable satisfying $\mathbb{E}(\epsilon_i | \boldsymbol{X}_i) = 0$ and $\mathbb{E}(\epsilon_i^2 | \boldsymbol{X}_i) \le \sigma_0^2$ for some $\sigma_0 > 0$. We consider matrix recovery in high dimensions, that is, $d_1 \times d_2$ can be much larger than the sample size n, making the problem ill-posed. It has been a common practice to assume that $\mathbf{\Theta}^*$ is (approximately) low-rank, and the resulting problem is referred to as low-rank matrix recovery.

The problem of low-rank matrix recovery can be naturally formulated as a nonconvex empirical risk minimization problem subject to a rank constraint. To find local optima of such a rank-constrained program, commonly used methods are Riemannian gradient descent [41] and Burer-Monteiro type gradient descent [2,7,25,38]. The former views the set of rank-r matrices as a smooth manifold, while the latter relies on the matrix factorization $\mathbf{\Theta} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{d_2 \times r}$, and $r = \operatorname{rank}(\mathbf{\Theta}^*)$ is assumed to be known. To relax the restrictive assumption that the true rank r is known *a priori*, [22]

and [43] studied the gradient method for solving the reparameterized program in an over-parameterized regime where $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r'}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2 \times r'}$, and $r' \geq r$ is an upper bound of the true rank.

Another line of research resorts to convex relaxation in order to obtain computationally feasible solutions. Similar to Lasso [37] in the context of sparse linear regression, convex low-rank matrix recovery methods are based on either constrained nuclear norm minimization or nuclear norm penalized least squares formulation. The nuclear norm of a matrix is defined as the sum of its singular values, and thus serves as a convex surrogate for its rank. We refer to [4–6,28–31] and [18] for an unavoidably incomplete list of notable works on exact and noisy low-rank matrix recovery through convex relaxation. In the context of multitask learning, [23] introduced an approach that utilizes the group Lasso penalty when only a small number of rows in the matrix $\boldsymbol{\Theta}^*$ are nonzero.

All the aforementioned methods, convex or nonconvex, are studied either in the noiseless setting or under a sub-Gaussian/sub-exponential assumption on the random error. However, both convex and nonconvex least squares estimators exhibit sub-optimal deviation bounds in the presence of heavy-tailed errors that only have a small number of finite moments. To make the estimator less sensitive to heavy-tailedness, a natural idea is to replace the ℓ_2-loss with a robust loss function, such as the ℓ_1-loss or the Huber loss [16]. For example, [10] proposed and studied nuclear norm penalized estimators using both the ℓ_1-loss and the Huber loss; [35] considered robust sparse reduced rank regression by minimizing the empirical Huber loss plus a combination of the nuclear norm and entry-wise ℓ_1-norm penalties. For methods that rely on nonconvex optimization with robust losses, [33] proposed a Riemannian sub-gradient method and proved the statistical properties of the iterates; [40] studied the statistical properties of vanilla gradient descent iterates for solving reparameterized (regularized) Huber loss minimization. In an over-parameterized regime, [26] showed that sub-gradient descent with the ℓ_1-loss function converges to the ground truth at a near-linear rate in the presence of arbitrarily large outliers.

In this paper, we propose a robust approach to recover an approximately low-rank matrix in a trace regression model with heavy-tailed and asymmetric error, which complements the extant literature on low-rank matrix recovery via convex relaxation. Borrowing ideas from robust (sparse) linear regressions [11,34], we adopt the Huber loss function with a diverging robustification parameter to achieve sub-Gaussian-type concentration bounds. We focus on three archetypal examples in matrix recovery: matrix compressed sensing, matrix completion and multitask regression. For each problem, we study the nonasymptotic deviation bounds of the nuclear norm penalized Huber estimator under both the Frobenius and nuclear norms, which match the minimax optimal rates. Our main contributions are as follows. First and foremost, we provide a comprehensive analysis of the nuclear norm penalized Huber regression estimator to gain robustness without compromising statistical efficiency. Our results either improve or complement those in [10,13] and [35]. For example, [10] considered robust matrix completion under symmetric error distribution and also required a constant lower bound for the error density function. [35] examined the Huber-type estimator for sparse multitask regression but their analysis cannot be directly extended to the non-sparse setting. Secondly, we provide a unified algorithmic framework, which is a matrix variant of the local adaptive majorize-minimization (LAMM) algorithm [12], to solve the three problems (matrix sensing, matrix completion and multitask regression) all at once. By constructing an isotropic quadratic function that locally majorizes the empirical Huber loss, the solution to each proximal optimization problem has a closed form, which considerably facilitates the implementation. Compared to many other algorithms used in the literature, our algorithm is first-order and thus more scalable to large data sets.

1.1. Related work and paper organization

Our model setting is closely related to that in [13], but the proposed robust estimators provably achieve sharper convergence rates than those obtained in [13]. More specifically, [13] proposed a two-step procedure, which in step one applies shrinkage operators to the empirical average $(1/n)\sum_{i=1}^{n} y_i \boldsymbol{X}_i$. The truncation level on y_i's, which appears in the final convergence rate, depends on the variance of y_i and thus is not proportional to the noise scale. In contrast, by employing the adaptive Huber loss as in [34] and the localized analysis developed by [12], we show that the convergence rates of our estimators are proportional to the noise scale for matrix sensing and multitask regression; see Theorem 3.3, Theorem 3.5 and the subsequent remarks for details. Moreover, [13] required ϵ_i to have bounded $(2k)$-th moment for some $k > 1$, while our estimators enjoy optimal rates as long as ϵ_i's have bounded variances. On the computational aspect, compared to the contractive Peaceman-Rachford splitting method and the alternating direction method of multiplier (ADMM) employed by [13] to solve the nuclear norm penalized programs in step two, the proposed matrix variant of the LAMM algorithm is first-order and has a lower computational cost per iteration. See Section 2.2 for a more detailed comparison of computational complexity.

Although the proposed estimators satisfy exponential deviation bounds when the noise distribution is asymmetric and has finite variance, this advantage is accompanied by a trade-off: they sacrifice a considerable level of robustness when facing adversarial contamination of the data. This is due to the use of a robustification parameter that increases with the sample size. In the case of adversarial contamination, recent studies have introduced robust estimators that showcase resistance to a small proportion of arbitrary outliers. For example, in sparse linear regression with Gaussian errors and adversarially corrupted labels, [9] demonstrated that the ℓ_1-penalized Huber's M-estimator attains the optimal rate of convergence, up to a logarithmic factor. Moreover, several recent studies [8,19,21,36] have specifically tackled the challenge of arbitrary outliers in the context of matrix sensing and matrix completion. For multitask regression, a robust multitask (reduced-rank) regression approach was introduced by [32] for simultaneous modeling and outlier detection. To address data contamination caused by arbitrary outliers, they formulated the problem as a regularized multivariate regression with a sparse mean-shift parametrization and developed a thresholding-based iterative procedure for optimization. It is worth noting that our methods and theory diverge from the conventional notion of robust statistics. While the aforementioned works assume sub-Gaussian or Gaussian noises, our work places emphasis on the distinct assumption of heavy-tailed errors rather than corruption by (arbitrary) outliers. The proposed methods and analysis therefore provide a useful complement to the current body of research on robust matrix completion and reduced-rank regression.

The rest of the paper proceeds as follows. In Section 2, we first review the trace regression model with three prototypical applications. Next, we introduce the nuclear norm penalized robust matrix estimator via the use of adaptive Huber loss, followed by a unified algorithm that applies to all three settings. We provide non-asymptotic high probability bounds for the proposed estimators case-by-case in Section 3. Section 4 presents our numerical experiments, conducted to demonstrate the advantage of our methods over their non-robust counterparts and to corroborate the theoretical findings that the convergence rates are proportional to the noise scale under the matrix sensing and multitask regression settings. All the proofs are relegated to the Appendix in the Supplementary Material [42].

Notation. For a matrix $\boldsymbol{A} = (A_{jk})_{1\le j\le d_1, 1\le k\le d_2} \in \mathbb{R}^{d_1\times d_2}$, its singular values are denoted as $\sigma_1(\boldsymbol{A}) \ge \sigma_2(\boldsymbol{A}) \ge \cdots \ge \sigma_{\min(d_1,d_2)}(\boldsymbol{A})$. Define its operator norm $\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A})$, its Frobenius norm $\|\boldsymbol{A}\|_{\mathrm{F}} = \sum_{j=1}^{\min(d_1,d_2)} \sigma_j^2(\boldsymbol{A})$, its nuclear norm $\|\boldsymbol{A}\|_* = \sum_{j=1}^{\min(d_1,d_2)} \sigma_j(\boldsymbol{A})$ and its max norm $\|\boldsymbol{A}\|_\infty = \max_{1\le j\le d_1}\max_{1\le k\le d_2}|A_{jk}|$. For two matrices $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{d_1\times d_2}$, let $\langle \boldsymbol{A},\boldsymbol{B}\rangle$ be the matrix inner product

defined as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B})$. We use $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{R}^{d_1 d_2}$ to denote the long vector obtained by stacking the columns of $\boldsymbol{A}$.

2. Robust matrix recovery via adaptive Huber loss

2.1. Model and methods

Suppose we have collected n i.i.d. data points $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ generated according to the following heteroscedastic trace regression model

$$y_i = \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle + \epsilon_i, \tag{2.1}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$'s are random measurement matrices, and ϵ_i's are additive random noise variables satisfying $\mathbb{E}(\epsilon_i | \boldsymbol{X}_i) = 0$ and $\mathbb{E}(\epsilon_i^2 | \boldsymbol{X}_i) \le \sigma_0^2$. Based on the noisy observations $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$, we are interested in recovering the unknown matrix $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ that is either exactly or approximately low-rank. More specifically, assume for some $0 \le q \le 1$ and $\rho > 0$ that

$$\boldsymbol{\Theta}^* \in \mathcal{B}_q(\rho) := \left\{ \boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \sum_{j=1}^{\min(d_1, d_2)} \sigma_j(\boldsymbol{\Theta})^q \le \rho \right\}. \tag{2.2}$$

In particular, $\mathcal{B}_0(\rho) = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \mathrm{rank}(\boldsymbol{\Theta}) \le \rho\}$ denotes the set of matrices with rank at most ρ, and $\mathcal{B}_q(\rho)$ with $0 < q \le 1$ is set of approximately low-rank matrices. Throughout the rest of the paper, we assume without loss of generality that $d_1 \ge d_2$.

The difficulty of recovering $\boldsymbol{\Theta}^*$ varies depending on the random structures of the measurement matrices $\boldsymbol{X}_i$. Below we list three prototypical applications of model (2.1), which will be the main focus of this work.

(i) *Matrix sensing*: Matrix sensing often assumes that the entries of $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ are independently generated from the $\mathcal{N}(0,1)$ distribution. More generally, $\mathrm{vec}(\boldsymbol{X}_i)$'s are assumed to be zero-mean sub-Gaussian/sub-exponential random vectors.

(ii) *Matrix completion*: In matrix completion, $\boldsymbol{X}_i$ are randomly drawn from the set

$$\mathcal{X} = \{\boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2), 1 \le j \le d_1, 1 \le k \le d_2\},$$

where $\boldsymbol{e}_1(d), \ldots, \boldsymbol{e}_d(d)$ are the canonical basis vectors in $\mathbb{R}^d$.

(iii) *Multitask regression*: The multitask (reduced-rank) regression assumes

$$\boldsymbol{y}_i = (\boldsymbol{\Theta}^*)^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{2.3}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{id_2})^{\mathrm{T}} \in \mathbb{R}^{d_2}$ are observed response vectors, $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ are covariate vectors, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is the target regression coefficient matrix, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{id_2})^{\mathrm{T}} \in \mathbb{R}^{d_2}$ are independent zero-mean random noise vectors. For $i = 1, \ldots, n$ and $k = 1, \ldots, d_2$, define

$$y_{(i-1)d_2+k} = y_{ik}, \quad \boldsymbol{X}_{(i-1)d_2+k} = \boldsymbol{x}_i \boldsymbol{e}_k^{\mathrm{T}}(d_2) \text{ and } \epsilon_{(i-1)d_2+k} = \epsilon_{ik}. \tag{2.4}$$

Then the sample $\{(y_j, \boldsymbol{X}_j)\}_{j=1}^N$ with $N = nd_2$ satisfies model (2.1).

For matrix sensing and matrix completion with noisy measurements, a popular approach is the the following convex relaxation approach [4]

$$\widehat{\boldsymbol{\Theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in C}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta}\rangle)^2 + \lambda\|\boldsymbol{\Theta}\|_*\right\}, \tag{2.5}$$

where C is a convex feasible set of $\mathbb{R}^{d_1\times d_2}$ and $\lambda > 0$ is a regularization parameter. When $C = \mathbb{R}^{d_1\times d_2}$, $\widehat{\boldsymbol{\Theta}}_\lambda$ is the matrix analog of the Lasso estimator for linear regression [37]. The statistical properties of $\widehat{\boldsymbol{\Theta}}_\lambda$ in (2.5), mainly nonasymptotic deviation bounds under various matrix norms, have been studied in the literature when the additive noises ϵ_i are either Gaussian or sub-Gaussian. The performance of such a least-square-type estimator may break down quickly when the noise distribution is heavier-tailed. This is because outliers occur more frequently and the square loss is very sensitive to outliers. The impact of heavy-tailed errors on low-rank matrix recovery can be alleviated by replacing the ℓ_2-loss with a more robust loss function, typified by the ℓ_1-loss and the Huber loss [10]. When the error distribution is not only heavy-tailed but also asymmetric around zero, the use of ℓ_1-loss or Huber loss with a fixed tuning parameter induces a bias that remains non-negligible as the number of measurements grows. For a better trade-off between robustness and bias, in the following we propose to use adaptive Huber loss [11,34] for robust low-rank matrix recovery, with a focus on the above three prototypical applications.

For matrix sensing and completion problems, i.e. applications (i) and (ii), we define the empirical loss function to be

$$\widehat{L}_\tau(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}\ell_\tau(y_i - \langle \boldsymbol{X}_i, \boldsymbol{\Theta}\rangle), \quad \boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}, \tag{2.6}$$

where $\ell_\tau(u) = \min\{u^2/2, \tau|u| - \tau^2/2\}$ denotes the adaptive Huber loss parameterized by $\tau = \tau_n > 0$, referred to as the robustification parameter in [34]. For any pre-specified convex subset C of $\mathbb{R}^{d_1\times d_2}$, we consider the following nuclear norm penalized robust regression estimator

$$\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in C}\left\{\widehat{L}_\tau(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_*\right\}, \tag{2.7}$$

where $\tau = \tau_n > 0$ and $\lambda = \lambda_n > 0$ are the robustification and regularization parameters respectively.

For the multitask regression problem – Application (iii), recall that the vector-valued observations $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ can be written as $\{(y_j, \boldsymbol{X}_j)\}_{j=1}^N$ ($N = nd_2$) via (2.4) so model (2.1) can be used. The classical reduced-rank regression method is based on solving the rank-constrained problem [17]

$$\min_{\operatorname{rank}(\boldsymbol{\Theta})\le r}\left\{\sum_{i=1}^{n}\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2^2 = \sum_{j=1}^{N}(y_j - \langle \boldsymbol{X}_j, \boldsymbol{\Theta}\rangle)^2\right\},$$

for which an analytic solution is available. To robustify this classical procedure, similarly to the formulation (2.7) one may naively apply the Huber loss to each residual $y_j - \langle \boldsymbol{X}_j, \boldsymbol{\Theta}\rangle$. This, however, is no longer plausible because $\boldsymbol{X}_j$'s are now dependent random matrices. Moreover, since we do not impose independence on the entries of $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{id_2})^{\mathrm{T}}$, ϵ_j's defined in (2.4) may also be highly correlated. We propose to replace the ℓ_2-loss on $\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2$ with the Huber loss, leading to $\min_{\operatorname{rank}(\boldsymbol{\Theta})\le r}\sum_{i=1}^n \ell_\tau(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2)$, which is a highly nonconvex problem. Similarly to (2.7), we resort to convex relaxation and consider the following nuclear norm penalized estimator

$$\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell_\tau\big(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2\big) + \lambda\|\boldsymbol{\Theta}\|_*\right\}. \tag{2.8}$$

In Section 3, we characterize the nonasymptotic statistical accuracy for the robust low-rank estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ defined in (2.7) and (2.8) when the noise variables only have bounded variances. The key is to seek a suitable choice of τ and λ jointly to trade off among robustness, bias and approximation errors.

2.2. Algorithms

To solve the nuclear norm penalized optimization programs (2.7) and (2.8), in this section we present a unified algorithm by extending the local adaptive majorize-minimization (LAMM) principle proposed in [12] to matrix settings. Recall that the proposed nuclear norm penalized Huber regression estimators have a general form

$$\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in C}\{\widehat{L}_\tau(\boldsymbol{\Theta}) + \lambda\|\boldsymbol{\Theta}\|_*\},$$

where $\widehat{L}_\tau(\boldsymbol{\Theta})$ is the empirical loss in (2.6) or (2.8), and C is taken to be $\mathbb{R}^{d_1\times d_2}$ or $\{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}: \|\boldsymbol{\Theta}\|_\infty \le \alpha_0\}$ for some $\alpha_0 > 0$. The main idea of the LAMM principle is to construct an isotropic quadratic function that locally majorizes the objective function at each iteration. In the matrix setting, given the previous iterate $\boldsymbol{\Theta}^{(k-1)}$ at the k-th iteration, define the quadratic function

$$F(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)},\phi_k) = \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}) + \langle\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}),\boldsymbol{\Theta}-\boldsymbol{\Theta}^{(k-1)}\rangle + \frac{\phi_k}{2}\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^{(k-1)}\|_{\mathrm{F}}^2,$$

satisfying $F(\boldsymbol{\Theta}^{(k-1)};\boldsymbol{\Theta}^{(k-1)},\phi_k) = \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)})$, where $\phi_k > 0$ is a quadratic parameter. Next, define the k-th iterate as

$$\boldsymbol{\Theta}^{(k)} \in \operatorname*{argmin}_{\boldsymbol{\Theta}\in C}\{F(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)},\phi_k) + \lambda\|\boldsymbol{\Theta}\|_*\}. \tag{2.9}$$

The parameter ϕ_k needs to be sufficiently large so that $\widehat{L}_\tau(\boldsymbol{\Theta}^{(k)}) \le F(\boldsymbol{\Theta}^{(k)};\boldsymbol{\Theta}^{(k-1)},\phi_k)$, which further implies

$$\begin{aligned}\widehat{L}(\boldsymbol{\Theta}^{(k)}) + \lambda\|\boldsymbol{\Theta}^{(k)}\|_* &\le F(\boldsymbol{\Theta}^{(k)};\boldsymbol{\Theta}^{(k-1)},\phi_k) + \lambda\|\boldsymbol{\Theta}^{(k)}\|_* \\ &\le F(\boldsymbol{\Theta}^{(k-1)};\boldsymbol{\Theta}^{(k-1)},\phi_k) + \lambda\|\boldsymbol{\Theta}^{(k-1)}\|_* \\ &= \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}) + \lambda\|\boldsymbol{\Theta}^{(k-1)}\|_*,\end{aligned}$$

where the second inequality is due to the optimality of $\boldsymbol{\Theta}^{(k)}$. This ensures the descent of the objective function at each iteration. To choose a sufficiently large ϕ_k, we start from a small value, say $\phi_0 = 0.01$, and inflate it by a factor $\gamma > 1$, say $\gamma = 2$, until the local majorization requirement $\widehat{L}_\tau(\boldsymbol{\Theta}^{(k)}) \le F(\boldsymbol{\Theta}^{(k)};\boldsymbol{\Theta}^{(k-1)},\phi_k)$ is met. Since $F(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)},\phi_k) \ge \widehat{L}_\tau(\boldsymbol{\Theta})$ when ϕ_k is no less than the largest eigenvalue of $\nabla^2\widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)})$, the iteration will stop after sufficiently many steps. Repeat the above steps until convergence (e.g., $\|\boldsymbol{\Theta}^{(k)}-\boldsymbol{\Theta}^{(k-1)}\|_{\mathrm{F}} \le \epsilon$ for a sufficiently small $\epsilon > 0$) or until the maximum number of iterations is reached.

The main benefit of minimizing a penalized isotropic quadratic objective function is that the minimizer often has a closed form. For any matrix $\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}$ with rank r, consider the singular value decomposition (SVD) $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are, respectively, $d_1\times r$ and $d_2\times r$ matrices with orthonormal columns, and $\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_i\}_{1\le i\le r})$ is an $r\times r$ diagonal matrix with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$.

Algorithm 1 LAMM algorithm for regularized adaptive Huber trace regression

Algorithm: $\{\boldsymbol{\Theta}^{(k)}\}_{k=1}^{\infty} \leftarrow \mathrm{LAMM}(\lambda, \boldsymbol{\Theta}^{(0)}, \phi_0, \gamma, \epsilon)$

Input: $\lambda, \boldsymbol{\Theta}^{(0)}, \phi_0, \gamma, \epsilon$

for $k = 1, 2, \ldots$ until $\|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^{(k-1)}\|_2 \le \epsilon$ **do**
  repeat
    $\boldsymbol{\Theta}^{(k)} \leftarrow S(\boldsymbol{\Theta}^{(k-1)} - \phi_k^{-1} \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}), \phi_k^{-1}\lambda)$
    $\boldsymbol{\Theta}^{(k)} \leftarrow \Pi_C(\boldsymbol{\Theta}^{(k)})$
    if $F(\boldsymbol{\Theta}^{(k)}; \boldsymbol{\Theta}^{(k-1)}, \phi_k) < \widehat{L}_\tau(\boldsymbol{\Theta}^{(k)})$ **then**
      $\phi_k \leftarrow \gamma \cdot \phi_k$
    end if
  until $F(\boldsymbol{\Theta}^{(k)}; \phi_k, \boldsymbol{\Theta}^{(k-1)}) \ge \widehat{L}_\tau(\boldsymbol{\Theta}^{(k)})$
  return $\boldsymbol{\Theta}^{(k)}$
end for

Output: $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(T)}$

For $\lambda > 0$, define the soft-thresholding operator $S(\boldsymbol{\Theta}, \lambda) = \boldsymbol{U} \cdot \mathrm{diag}(\{\max(\sigma_i - \lambda, 0)\}_{1 \le i \le r}) \cdot \boldsymbol{V}^{\mathrm{T}}$. By Theorem 2.1 in [3], $\boldsymbol{\Theta}^{(k)}$ given in (2.9) with $C = \mathbb{R}^{d_1 \times d_2}$ admits the closed-form expression

$$\boldsymbol{\Theta}^{(k)} = T_{\lambda,\phi_k}(\boldsymbol{\Theta}^{(k-1)}) := S(\boldsymbol{\Theta}^{(k-1)} - \phi_k^{-1} \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)}), \phi_k^{-1}\lambda).$$

For a general convex subset $C \subseteq \mathbb{R}^{d_1 \times d_2}$, we can update $\boldsymbol{\Theta}^{(k)}$ as

$$\boldsymbol{\Theta}^{(k)} = \Pi_C(T_{\lambda,\phi_k}(\boldsymbol{\Theta}^{(k-1)})),$$

where Π_C denotes Euclidean projection onto the subspace C. When $C = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Theta}\|_\infty \le \alpha_0\}$, for example, $\Pi_C(\boldsymbol{\Theta}) = (\max\{\min(\boldsymbol{\Theta}_{jk}, \alpha_0), -\alpha_0\})_{1 \le j \le d_1, 1 \le k \le d_2}$. We summarize the key steps in Algorithm 1.

As a unified algorithm, Algorithm 1 applies to all three problems considered in this paper, matrix sensing, matrix completion and multitask regression. In terms of complexity, at each iteration $\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^{(k-1)})$ and $T_{\lambda,\phi_k}(\boldsymbol{\Theta}^{(k-1)})$ can be computed in $O(nd_1d_2)$ and $O(d_1d_2^2)$ operations (assuming $d_1 \ge d_2$), respectively [39]. On the other hand, [13] employed the contractive Peaceman-Rachford splitting method for matrix sensing and multitask regression, and an ADMM-based algorithm for matrix completion. In addition to the operations described above, each ADMM iterate also involves computing the inverse of $2\mathbb{X}^{\mathrm{T}}\mathbb{X}/n + \mathbf{I}_{d_1d_2}$, where $\mathbb{X}$ is an $n \times d_1d_2$ matrix whose i-th row is $\mathrm{vec}(\boldsymbol{X}_i)$, and $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. By applying the Sherman–Morrison–Woodbury formula, this step can be implemented in $O(\min\{n, d_1d_2\}^3)$ operations. Still, the computational complexity and storage cost (per iteration) of ADMM are much higher than the LAMM algorithm in the context of matrix completion, especially for large-scale datasets.

3. Theoretical guarantees

In this section, we establish the finite-sample statistical properties of the robust estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ for matrix sensing, matrix completion and multitask regression. Throughout, the noise variables ϵ_i in (2.1) and $\boldsymbol{\epsilon}_i$ in (2.3) are assumed to have bounded variance only, and we do not require independence between ϵ_i and $\boldsymbol{X}_i$ or $\boldsymbol{\epsilon}_i$ and $\boldsymbol{x}_i$.

3.1. Matrix sensing

In the case of matrix compressed sensing, we set $\mathcal{C} = \mathbb{R}^{d_1 \times d_2}$ in (2.7), and impose the following assumptions.

(A1) $\boldsymbol{\Theta}^* \in \mathcal{B}_q(\rho)$ for some $0 \le q \le 1$ and $\rho > 0$.

(A2) The random matrix $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ satisfies (i) $\mathbb{E}\boldsymbol{X}_i = \boldsymbol{0}$, and (ii) $\mathrm{vec}(\boldsymbol{X}_i) \in \mathbb{R}^{d_1 d_2}$ is sub-exponential, that is, there exists a constant $\nu_0 \ge 1$ such that for any $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$ and $u \ge 0$,

$$\mathbb{P}\big(|\mathrm{vec}(\boldsymbol{A})^{\mathrm{T}}\mathrm{vec}(\boldsymbol{X}_i)| \ge \nu_0 \|\boldsymbol{A}\|_{\mathrm{F}} \cdot u\big) \le 2e^{-u}.$$

Moreover, there exists a constant $c_l > 0$ such that $\lambda_{\min}\big(\mathbb{E}\mathrm{vec}(\boldsymbol{X}_i)\mathrm{vec}(\boldsymbol{X}_i)^{\mathrm{T}}\big) \ge c_l$.

(A3) The noise variables ϵ_i are such that $\mathbb{E}(\epsilon_i|\boldsymbol{X}_i) = 0$ and $\mathbb{E}(\epsilon_i^2|\boldsymbol{X}_i) \le \sigma_0^2$ (almost surely) for some constant $\sigma_0 > 0$.

Remark 1. The parameter ν_0 is often referred to as the sub-exponential parameter. For various well-behaved distributions on $\mathbb{R}^{d_1 \times d_2}$, the associated sub-exponential parameters are independent of the dimensions d_1 and d_2. As prototypical examples, the distributions listed below satisfy Condition (A2) with dimension-free parameters ν_0 and c_l.

(i) (Multivariate normal) $\mathrm{vec}(\boldsymbol{X}_i)$ follow s $\mathcal{N}(\boldsymbol{0}, \Sigma)$ with a positive-definite $\Sigma \in \mathbb{R}^{(d_1 d_2)\times(d_1 d_2)}$.

(ii) (Uniform distribution on the Euclidean sphere) $\boldsymbol{X}_i$ follows the uniform distribution on the sphere centered at the origin with radius $(d_1 d_2)^{1/2}$, namely, $\{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{X}\|_{\mathrm{F}} = (d_1 d_2)^{1/2}\}$.

(iii) (Uniform distribution on the ℓ_1-ball) $\boldsymbol{X}_i$ follows the uniform distribution on the ℓ_1-norm ball centered at the origin with radius $r \asymp d_1 d_2$, that is, $\{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} : \sum_{j=1}^{d_1}\sum_{k=1}^{d_2} |X_{jk}| \le r\}$.

Here, we note that the multivariate distributions in (i) and (ii) are not only sub-exponential but also sub-Gaussian.

To derive the convergence rate of $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ under either the Frobenius norm or the nuclear norm, we first define a probability event that concerns the local restricted strong convexity (RSC) of the empirical loss $\widehat{L}_\tau(\cdot)$. For $s, l > 0$, define the Frobenius norm ball and trace norm cone

$$\mathbb{B}(s) = \{\boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Delta}\|_{\mathrm{F}} \le s\} \quad \text{and} \quad \mathbb{C}(l) = \{\boldsymbol{\Delta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Delta}\|_* \le l\|\boldsymbol{\Delta}\|_{\mathrm{F}}\}, \tag{3.1}$$

respectively.

Definition 3.1 (Local restricted strong convexity). Given radius parameters $s, l > 0$ and a curvature parameter $\kappa > 0$, define the event

$$\mathcal{E}(s, l, \kappa) = \left\{ \inf_{\boldsymbol{\Theta} \in \boldsymbol{\Theta}^* + \mathbb{B}(s) \cap \mathbb{C}(l)} \frac{\langle \nabla \widehat{L}_\tau(\boldsymbol{\Theta}) - \nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^* \rangle}{\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2} \ge \kappa \right\}, \tag{3.2}$$

which concerns the local restricted strong convexity of the empirical loss function.

We first provide a deterministic result on the convergence rate of $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$: for any choice of λ such that $\|\nabla \widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \le \lambda/2$, and conditioned on event $\mathcal{E}(s, l, \kappa)$ with suitably chosen (s, l), we are guaranteed that

$$\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \lesssim \sqrt{\rho}\,(\lambda/\kappa)^{1-q/2} \quad \text{and} \quad \|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_* \lesssim \rho(\lambda/\kappa)^{1-q}.$$

Theorem 3.2. *Assume Condition (A1) holds. Let (λ, s, l, κ) satisfy $\lambda \geq 2\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$,*

$$s \geq 9.15\rho(\lambda/\kappa)^{1-q/2} \quad \text{and} \quad l \geq 6.1\rho(\kappa/\lambda)^{q/2}. \tag{3.3}$$

Conditioned on the event $\mathcal{E}(s, l, \kappa)$, the error matrix $\widehat{\mathbf{\Delta}} := \widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*$ satisfies

$$\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} \leq 9.15\sqrt{\rho}\left(\frac{\lambda}{\kappa}\right)^{1-q/2} \quad \text{and} \quad \|\widehat{\mathbf{\Delta}}\|_* \leq 56\rho\left(\frac{\lambda}{\kappa}\right)^{1-q}.$$

In the following two propositions, we first derive an upper bound of $\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$, and then establish the local RSC property of the empirical Huber loss function $\widehat{L}_\tau(\cdot)$. Together, these results show that with properly chosen λ, τ that depend on (n, d, s, l) along with the distributional parameters in Conditions (A2) and (A3), the event $\{\lambda \geq 2\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2\} \cap \mathcal{E}(s, l, c_l/4)$ occurs with high probability.

Proposition 3.1. *Assume Conditions (A2) and (A3) hold. For any $\sigma \geq \sigma_0$ and $z > 0$, the empirical Huber loss $\widehat{L}_\tau(\cdot)$ with $\tau = \sigma\sqrt{n/(3d+z)}$ satisfies*

$$\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \leq 10\nu_0 \cdot \sigma\sqrt{\frac{3d+z}{n}} \tag{3.4}$$

with probability at least $1 - e^{-z}$, where $d = d_1 + d_2$.

Proposition 3.2. *Assume Conditions (A2) and (A3) hold. For any $s, l > 0$ and $z > 0$, let τ and n satisfy*

$$\tau \geq 4\nu_0\sqrt{(2\sigma_0^2 + 96\nu_0^2 s^2)/c_l} \quad \text{and} \quad n \geq C_1(\tau/s)^2(l^2 d + z), \tag{3.5}$$

where $d = d_1 + d_2$ and $C_1 > 0$ is a constant depending only on ν_0 and c_l. Then, the local RSC event $\mathcal{E}(s, l, \kappa)$ with $\kappa = c_l/4$ occurs with probability at least $1 - e^{-z}$.

Combining these high probability bounds with Theorem 3.2 leads to the convergence rate of $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$, as stated in the following theorem.

Theorem 3.3. *Assume Conditions (A1)–(A3) hold. For any $z > 0$, the robust (approximately) low-rank matrix estimator $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ defined in (2.7) with $C = \mathbb{R}^{d_1 \times d_2}$, $\tau \asymp \sigma_0\sqrt{n/(d+z)}$ and $\lambda \asymp \sigma_0\sqrt{(d+z)/n}$ satisfies*

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0^{1-q/2}\sqrt{\rho}\left(\frac{d+z}{n}\right)^{1/2-q/4} \quad \text{and} \quad \|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_* \lesssim \sigma_0^{1-q}\rho\left(\frac{d+z}{n}\right)^{(1-q)/2}$$

with probability at least $1 - 2e^{-z}$ as long as $n \gtrsim \max\left\{(\rho/\sigma_0^q)^{2/(2-q)}, 1\right\}(d+z)$, where $d = d_1 + d_2$.

Remark 2. In the exact low-rank case, i.e. $q = 0$ and $\rho = r = \mathrm{rank}(\mathbf{\Theta}^*)$, the results in Theorem 3.3 imply that with high probability (over both the random sensing matrices $\boldsymbol{X}_i$ and noise variables ϵ_i), the robust estimator $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ satisfies with high probability that

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0\sqrt{\frac{rd}{n}} \ \text{ as long as } \ n \gtrsim rd.$$

Within a constant independent of (n, r, d_1, d_2) and noise scale σ_0, this upper bound matches the information-theoretic lower bound established by [5] when ϵ_i are i.i.d. $\mathcal{N}(0, \sigma_0^2)$. The robustness manifests in two aspects. First, the noise distribution is only required to have bounded variance as opposed to sub-Gaussian tails. Secondly, we assume the random vector $\mathrm{vec}(\boldsymbol{X}_i)$ is sub-exponential, whereas $\mathrm{vec}(\boldsymbol{X}_i)$ is often assumed to have i.i.d. Gaussian/sub-Gaussian entries in the literature.

Remark 3. For matrix compressed sensing, based on a shrinkage principle [13] also proposed a robust low-rank estimator, which achieves near-optimal rate under heavy-tailed noise distributions. Its recovery guarantees (see Theorem 3 therein), however, depend on more stringent assumptions as needed in Theorem 3.3. In addition to Conditions (A2) and (A3), [13] assumed further that (i) $\mathrm{vec}(\boldsymbol{X}_i)$ is sub-Gaussian, and (ii) $\mathbb{E}|y_i|^{2k} \le M_k$ for some $k > 1$. Under these conditions and in the exact low-rank case (for brevity), their truncate/shrinkage estimator, denoted by $\widetilde{\boldsymbol{\Theta}}$, satisfies with high probability the bound

$$\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \le M_k^{1/(2k)} \sqrt{\frac{rd}{n}} \quad \text{as long as} \quad n \gtrsim rd.$$

The above convergence rate is sub-optimal in terms of its dependence on the noise scale σ_0. As the noise scale decreases, $M_k^{1/(2k)}$ remains to be bounded away from zero because

$$M_k^{1/k} > \mathbb{E} y_i^2 = \mathbb{E}\langle \boldsymbol{X}_i, \boldsymbol{\Theta}^* \rangle^2 + \mathbb{E}(\epsilon_i^2).$$

Remark 4. The sample size requirement in Theorem 3.3 becomes more stringent as σ_0 goes to 0 when $q \neq 0$. This is an artifact of the technical argument used in the proof of the theorem. A similar sample size requirement, characterized by its inverse proportionality to the moment of the response, can be found in Theorem 3 of [13]. To modify the sample size requirement, we can choose

$$\tau \asymp \max(\sigma_0, 1)\sqrt{n/(d+z)} \quad \text{and} \quad \lambda \asymp \max(\sigma_0, 1)\sqrt{(d+z)/n}$$

for a given $z > 0$. This results in a revised sample size requirement of $n \gtrsim \max(\rho^{2/(2-q)}, 1)(d+z)$, accompanied by an error bound

$$\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}} \lesssim \max(\sigma_0^{1-q/2}, 1)\sqrt{\rho}\left(\frac{d+z}{n}\right)^{1/2-q/4}$$

with probability at least $1 - 2e^{-z}$. This error bound is similar to the deviation bound in Corollary 5 of [28], but it should be noted that they are not proportional to the noise level. Additionally, Theorem 3.5 in Section 3.3 also relies on a similar technical argument, necessitating a larger sample size as σ_0 approaches zero.

3.2. Matrix completion

This subsection investigates matrix completion under the following assumptions.

- (B1) $\boldsymbol{\Theta}^* \in \mathcal{B}_q(\rho)$ and $\|\boldsymbol{\Theta}^*\|_\infty \le \alpha_0$ for some $\alpha_0 > 0$. We thus set $C = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Theta}\|_\infty \le \alpha_0\}$ in (2.7) so that $\boldsymbol{\Theta}^* \in C$.
- (B2) $\boldsymbol{X}_i \in \mathbb{R}^{d_1 \times d_2}$ is uniformly sampled from $\{\boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\}_{1 \le j \le d_1, 1 \le k \le d_2}$, where $\{\boldsymbol{e}_j(d)\}_{j=1}^d$ are the canonical basis vectors in $\mathbb{R}^d$. Specifically, $\mathbb{P}\{\boldsymbol{X}_i = \boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\} = (d_1 d_2)^{-1}$.
- (B3) $\mathbb{E}(\epsilon_i | \boldsymbol{X}_i) = 0$ and $\mathbb{E}(\epsilon_i^2 | \boldsymbol{X}_i) \le \sigma_0^2$ (almost surely) for some constant $\sigma_0 > 0$.

Remark 5. In addition to the assumption that $\mathbf{\Theta}^*$ is of (approximately) low-rank, we require in Condition (B1) that $\|\mathbf{\Theta}^*\|_\infty \le \alpha_0$ for some $\alpha_0 > 0$. Past works on noisy matrix completion also imposed the same or similar conditions. For instance, [18] and [27] assumed that $\|\mathbf{\Theta}^*\|_\infty$ is bounded; [29] and [13] required the spikiness ratio $\|\mathbf{\Theta}^*\|_\infty / \|\mathbf{\Theta}^*\|_{\mathrm{F}}$ to be bounded; [4] and [6] relied on matrix incoherence conditions. Without such extra conditions, the number of measurements should satisfy $n \asymp d_1 d_2$ in order to recover $\mathbf{\Theta}^*$ in the worst case; see [6] and [29] for details.

Similarly to the matrix sensing case, the key steps to establish the convergence rate of $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ are (i) an upper bound of $\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2$ as shown in Proposition 3.3 below, and (ii) a lower bound for

$$\langle \nabla \widehat{L}_\tau(\mathbf{\Theta}) - \nabla \widehat{L}_\tau(\mathbf{\Theta}^*), \mathbf{\Theta} - \mathbf{\Theta}^* \rangle$$

uniformly over $\mathbf{\Theta}$ in a neighborhood of $\mathbf{\Theta}^*$. The sparsity of $\boldsymbol{X}_i$ in this case (see Condition (B2)) introduces more subtleties into the analysis of the latter, as we will see from Proposition 3.4.

Proposition 3.3. *Assume Conditions (B2) and (B3) hold. For any $\sigma \ge \sigma_0$ and $z > 0$, the loss function $\widehat{L}_\tau(\cdot)$ with $\tau = \sigma\sqrt{n/\{d_2(z+\log d)\}}$ satisfies with probability at least $1 - e^{-z}$ that*

$$\|\nabla \widehat{L}_\tau(\mathbf{\Theta}^*)\|_2 \le (3\sigma_0 + 2\sigma/3)\sqrt{\frac{z+\log d}{d_2 n}}, \tag{3.6}$$

where $d = d_1 + d_2$.

Proposition 3.4. *Assume Conditions (B2) and (B3) hold. For any $s, l > 0$ and $z > 0$, let τ and n satisfy*

$$\tau^2 \ge 16 \max[ns^2/\{l^2 d_1^2 d_2 (z+\log d)\}, \sigma_0^2] \quad \text{and} \quad n \ge d_2 \log d,$$

where $d = d_1 + d_2$. Define the constrain set

$$\mathbb{A}(s,l) = \left\{ \mathbf{\Delta} \in \mathbb{B}(s) \cap \mathbb{C}(l) : \frac{\|\mathbf{\Delta}\|_\infty^2}{\|\mathbf{\Delta}\|_{\mathrm{F}}^2/(d_1 d_2)} \le \frac{1}{8}\sqrt{\frac{n}{z+\log d}} \right\}, \tag{3.7}$$

where $\mathbb{B}(s)$ and $\mathbb{C}(l)$ are given in (3.1). Then, for all $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ with $\mathbf{\Delta} := \mathbf{\Theta} - \mathbf{\Theta}^* \in \mathbb{A}(s,l)$, we have with probability at least $1 - e^{-z}$ that

$$\langle \nabla \widehat{L}_\tau(\mathbf{\Theta}) - \nabla \widehat{L}_\tau(\mathbf{\Theta}^*), \mathbf{\Theta} - \mathbf{\Theta}^* \rangle \ge \frac{1}{4 d_1 d_2}\|\mathbf{\Delta}\|_{\mathrm{F}}^2 - C_0 l^2 \frac{d_1(z+\log d)}{n}\|\mathbf{\Delta}\|_\infty^2,$$

where $C_0 > 1$ is an absolute constant.

With the above preparations, we now state the statistical guarantees for matrix completion under heavy-tailed noise.

Theorem 3.4. *Assume Conditions (B1)–(B3) hold. For any $z > 0$, set*

$$\tau \asymp \sigma\sqrt{\frac{n}{d_2(z+\log d)}} \quad \text{and} \quad \lambda \asymp \sigma\sqrt{\frac{z+\log d}{d_2 n}},$$

where $\sigma = \max\{\sigma_0, \alpha_0\}$ *and* $d = d_1 + d_2$. *Then, the robust (approximately) low-rank matrix estimator* $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ *defined in* (2.7) *with* $C = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{\Theta}\|_\infty \le \alpha_0\}$ *satisfies*

$$\frac{1}{d_1 d_2}\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \lesssim \max\left[\sigma^{2-q}\rho\left\{\frac{d_1(z+\log d)}{n}\right\}^{1-q/2}, \alpha_0^2\sqrt{\frac{z+\log d}{n}}\right] \tag{3.8}$$

and

$$\frac{1}{\sqrt{d_1 d_2}}\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_* \lesssim \max\left[\sigma^{1-q}\rho\left\{\frac{d_1(z+\log d)}{n}\right\}^{\frac{1-q}{2}}, \alpha_0^{\frac{2(1-q)}{2-q}}\frac{\rho^{\frac{1}{2-q}}}{(d_1 d_2)^{\frac{q}{2(2-q)}}}\left(\frac{z+\log d}{n}\right)^{\frac{1-q}{2(2-q)}}\right]$$

with probability at least $1 - 2e^{-z}$ *whenever* $n \gtrsim d_2(z + \log d)$.

Remark 6. In the context of matrix completion, one is interested in recovering a large low-rank data matrix from a highly incomplete subset of its entries. A natural assumption is $n \le d_1 d_2$, which in turn implies $\sqrt{\log(d)/n} \le d_1 \log(d)/n$, where $d = d_1 + d_2$ and $d_1 \ge d_2$. Therefore, taking $z = \log n$, the maximum in (3.8) is often given by its first term. In the exact low-rank case, the general results in Theorem 3.4 imply that the proposed robust estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ with $\tau \asymp \sigma_0\sqrt{n/(d_2 \log d)}$ and $\lambda \asymp \sigma_0\sqrt{\log(d)/(d_2 n)}$ satisfies the bound

$$\frac{1}{d_1 d_2}\|\widehat{\boldsymbol{\Theta}}_{\tau,\lambda} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \lesssim \max\{\alpha_0^2, \sigma_0^2\}\frac{r d_1 \log d}{n} \tag{3.9}$$

with high probability as long as $n \gtrsim d_2 \log d$. Under our notations, Theorem 6 in [20] shows that when $\epsilon_i \sim \mathcal{N}(0, \sigma_0^2)$ is independent of $\boldsymbol{X}_i$, there exist absolute constants $\beta \in (0,1)$ and $c > 0$ such that

$$\inf_{\widehat{\boldsymbol{\Theta}}} \sup_{\mathrm{rank}(\boldsymbol{\Theta}^*) \le r, \|\boldsymbol{\Theta}^*\|_\infty \le \alpha_0} \mathbb{P}\left\{\frac{1}{d_1 d_2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 > c\min(\sigma_0^2, \alpha_0^2)\frac{r d_1}{n}\right\} \ge \beta,$$

where $\inf_{\widehat{\boldsymbol{\Theta}}}$ is the infimum over all estimators $\widehat{\boldsymbol{\Theta}} \in \mathbb{R}^{d_1 \times d_2}$. Therefore, the rate derived in Theorem 3.4 is minimax optimal up to a logarithmic factor and a trailing term.

Remark 7 (Comparison to existing work on robust (noisy) matrix completion). In the context of matrix completion with heavy-tailed noise, several robust estimators have been proposed and studied. [27] proposed a two-step method that computes a truncation-type matrix estimator, denoted by $\widetilde{\boldsymbol{\Theta}}$, in step one and then solves the nuclear norm penalized optimization $\|\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2/(d_1 d_2) + \lambda\|\boldsymbol{\Theta}\|_*$. In the exact low-rank case, this two-step estimator satisfies a high probability bound, which is similar to (3.8) with $q = 0$, when ϵ_i is independent of $\boldsymbol{X}_i$ and has bounded variance. The independence assumption can be removed by slightly modifying the proof in [27]. For matrix sensing and multitask regression, it is unclear whether such a two-step procedure will also lead to robust estimates that satisfy sharp error bounds proportional only to the noise scale. Concurrently, [13] considered a similar two-step estimator, but their theoretical result requires a slightly stronger moment condition, i.e. $\mathbb{E}\{\mathbb{E}(\epsilon_i^2|\boldsymbol{X}_i)^k\} \le M_k$ for some $k > 1$. Our proposal is more relevant to [10], who also used the Huber loss for matrix completion in the presence of heavy-tailed errors. Their results, however, depend on stronger assumptions on the error distribution. In addition to Conditions (B1) and (B2), they further assumed that (i) the distribution of ϵ_i is symmetric around 0, and (ii) there exists a constant $C_1 > 0$ such that the cumulative distribution function $F(\cdot)$ of ϵ_i satisfies

$$F(x+\tau) - F(x-\tau) \ge 1/C_1^2 \quad \text{for all } |x| \le 2\alpha_0 \text{ and } \tau \le 2\alpha_0.$$

Under theses conditions and in the exact low-rank case, [10] proved that the nuclear norm penalized Huber regression estimator, denoted by $\check{\boldsymbol{\Theta}}$, satisfies

$$\frac{1}{d_1 d_2}\|\check{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 = O_{\mathbb{P}}\left\{ \max(\alpha_0^2, \tau^2) C_1^4 \frac{r d_1 \log(d_1 + d_2)}{n} \right\}$$

under the sample size requirement $n \gtrsim d_2 \log(d_2) \log(d_1 + d_2)$.

3.3. Multitask regression

In this section, we establish the statistical properties of the robust low-rank multitask (reduced-rank) regression estimator $\widehat{\boldsymbol{\Theta}}_{\tau,\lambda}$ (2.8). With slight abuse of notation, we write

$$\widehat{L}_\tau(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^n \ell_\tau\big(\|\boldsymbol{y}_i - \boldsymbol{\Theta}^{\mathrm{T}}\boldsymbol{x}_i\|_2\big), \quad \boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}, \tag{3.10}$$

where $\tau > 0$ is the robustification parameter.

(C1) $\boldsymbol{\Theta}^* \in \mathcal{B}_q(\rho)$ for some $0 \le q \le 1$ and $\rho > 0$.

(C2) $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ are i.i.d. zero-mean sub-Gaussian vectors, that is, there exists a (dimension-free) constant $\nu_0 \ge 1$ such that

$$\mathbb{E}e^{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_i} \le e^{\nu_0^2\|\boldsymbol{u}\|_2^2/2}, \text{ valid for any } \boldsymbol{u} \in \mathbb{R}^{d_1}.$$

Moreover, there exists a constant $c_l > 0$ such that $\lambda_{\min}\big(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\big) \ge c_l$.

(C3) The noise vectors $\boldsymbol{\epsilon}_i \in \mathbb{R}^{d_2}$ are such that $\mathbb{E}(\boldsymbol{\epsilon}_i|\boldsymbol{X}_i) = \boldsymbol{0}$ and $\lambda_{\max}\big(\mathbb{E}(\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^{\mathrm{T}}|\boldsymbol{x}_i)\big) \le \sigma_0^2$ (almost surely).

Proposition 3.5. *Assume Conditions (C2) and (C3) hold. For any $\sigma \ge \sigma_0$ and $z > 0$, choose $\tau = \sigma\sqrt{n/(z + \log d)}$ with $d = d_1 + d_2$. Then, it holds with probability at least $1 - e^{-z}$ that*

$$\|\nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*)\|_2 \le C\nu_0\sigma\sqrt{\frac{d(z + \log d)}{n}},$$

where $C > 0$ is a universal constant.

Proposition 3.6. *Assume Conditions (C2) and (C3) hold, and let $s, \tau > 0$ and $z > 0$ satisfy*

$$\tau \ge \max\left\{4\sigma_0\sqrt{d_2}, 2\nu_0 s\sqrt{2d_1 + 3z + 3\log n}\right\}. \tag{3.11}$$

Then, with probability at least $1 - 2e^{-z}$,

$$\langle\nabla\widehat{L}_\tau(\boldsymbol{\Theta}) - \nabla\widehat{L}_\tau(\boldsymbol{\Theta}^*), \boldsymbol{\Theta} - \boldsymbol{\Theta}^*\rangle \ge \frac{c_l}{2}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2 \text{ for all } \boldsymbol{\Theta} \in \boldsymbol{\Theta}^* + \mathbb{B}(s),$$

provided that $n \gtrsim \nu_0^4 c_l^{-2}(d_1 + z)$.

Theorem 3.5. *Assume Conditions (C1)-(C3) hold. For any $z > 0$, the robust (approximately) low-rank matrix estimator $\widehat{\mathbf{\Theta}}_{\tau,\lambda}$ defined in (2.8) with $\tau \asymp \sigma_0\sqrt{n/(z+\log d)}$ and $\lambda \asymp \sigma_0\sqrt{d(z+\log d)/n}$ ($d = d_1 + d_2$) satisfies the bounds*

$$\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \lesssim \sigma_0^{1-q/2}\sqrt{\rho}\left\{\frac{d(z+\log d)}{n}\right\}^{\frac{1}{2}-\frac{q}{4}} \quad \text{and} \quad \|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_* \lesssim \sigma_0^{1-q}\rho\left\{\frac{d(z+\log d)}{n}\right\}^{\frac{1-q}{2}}$$

with probability at least $1 - 3e^{-z}$ as long as

$$n \gtrsim \max\{(\rho/\sigma_0^q)^{2/(4-q)}, 1\} \cdot (d + z + \log n)(z + \log d).$$

Remark 8. Again, in the exact low-rank case where $\rho = r = \mathrm{rank}(\mathbf{\Theta}^*)$ and $q = 0$, Theorem 3.5 shows that for an arbitrary accuracy $\varepsilon > 0$, we have $\|\widehat{\mathbf{\Theta}}_{\tau,\lambda} - \mathbf{\Theta}^*\|_{\mathrm{F}} \le \varepsilon$ with an overwhelming probability provided that the number of measurements satisfies

$$n \gtrsim \sigma_0^2 \frac{rd\log d}{\varepsilon^2}. \tag{3.12}$$

This result improves Theorem 5 in [13] in several aspects. Under the multitask regression model (2.3), they assumed that

$$\lambda_{\max}\big(\mathbb{E}(\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}})\big) \le R < \infty, \qquad \max_{1\le i\le n, 1\le k\le d_2} \mathbb{E}\big\{\mathbb{E}(\epsilon_{ik}^2|\boldsymbol{x}_i)^k\big\} \le M_k < \infty \ \text{ for some } k > 1,$$

and for each i, $\epsilon_{i1}, \ldots, \epsilon_{id_2}$ are pairwise (conditionally) independent given $\boldsymbol{x}_i$. In contrast, Condition (C3) only assumes bounded variances and allows arbitrary dependency between ϵ_{ik}'s. For an arbitrary accuracy $\varepsilon > 0$, the truncated/shrinkage matrix estimator $\widetilde{\mathbf{\Theta}}$ proposed by [13] satisfies $\|\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}} \le \varepsilon$ with high probability provided

$$n \gtrsim (R + M_k^{1/k})\frac{rd\log d}{\varepsilon^2}. \tag{3.13}$$

Here the term $R + M_k^{1/k}$ can be much larger than σ_0^2 in (3.12). More importantly, as the noise scale σ_0 decays, R stays away from zero because

$$R \ge \lambda_{\max}\big(\mathbb{E}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}\big) = \lambda_{\max}\big((\mathbf{\Theta}^*)^{\mathrm{T}}\mathbf{\Sigma}\mathbf{\Theta}^* + \mathbb{E}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\mathrm{T}}\big) \ge \lambda_{\max}\big((\mathbf{\Theta}^*)^{\mathrm{T}}\mathbf{\Sigma}\mathbf{\Theta}^*\big).$$

4. Numerical studies

4.1. Finite-sample performance

In this section, we perform simulation studies to assess the finite-sample performance of the nuclear norm penalized adaptive Huber trace regression method (Nuclear-AH) in all three problems. As a benchmark, we implement the nuclear norm penalized least squares (Nuclear-LS) estimator also via the LAMM algorithm.

In addition to the regularization parameter λ, the use of an adaptive Huber loss also involves a robustification parameter τ that changes with data scales. We set $\tau = c_\tau \cdot a_{n,d}$ and $\lambda = c_\lambda \cdot b_{n,d}$, where c_τ and c_λ are positive constants that are independent of (n,d) but depend on the noise scale, and $a_{n,d}$ and $b_{n,d}$ are determined by the theoretical results in Section 3, as follows:

Matrix Sensing	Normal	t	Pareto
Nuclear-LS	0.227 (0.010)	0.173 (0.052)	0.169 (0.093)
Nuclear-AH	0.227 (0.010)	0.132 (0.008)	0.107 (0.007)

Matrix Completion	Normal	t	Pareto
Nuclear-LS	0.424 (0.021)	0.280 (0.047)	0.315 (0.041)
Nuclear-AH	0.445 (0.022)	0.223 (0.023)	0.252 (0.022)

Multitask Regression	Normal	t	Pareto
Nuclear-LS	0.228 (0.005)	0.213 (0.112)	0.237 (0.181)
Nuclear-AH	0.228 (0.005)	0.148 (0.003)	0.120 (0.003)

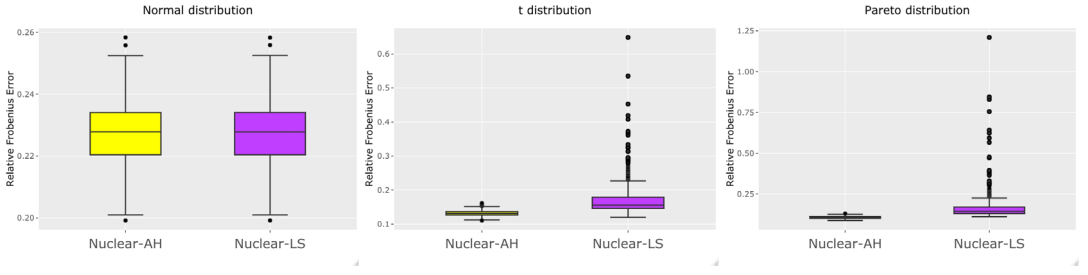
Table 1. Mean relative Frobenius error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}/\|\boldsymbol{\Theta}^*\|_{\mathrm{F}}$ (with standard deviations in parentheses), averaged over 500 replications, under the matrix sensing, matrix completion and multitask regression settings.

(a) For matrix sensing, we choose $a_{n,d} = \sqrt{n/d}$ and $b_{n,d} = \sqrt{d/n}$.
(b) For matrix completion, we choose $a_{n,d} = \sqrt{n/(d\log d)}$ and $b_{n,d} = \sqrt{\log(d)/(dn)}$.
(c) For multitask learning, we choose $a_{n,d} = \sqrt{n/\log d}$ and $b_{n,d} = \sqrt{d\log(d)/n}$.

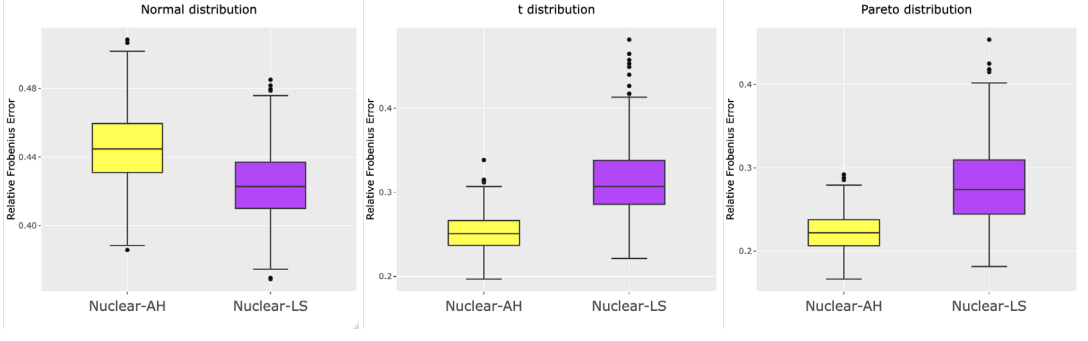
Then we follow the following steps to tune c_τ and c_λ.

(i) First, choose the constant c_λ in the Nuclear-LS method via five-fold CV with the absolute median loss as the criterion. In particular, we use the "one-standard-error" rule, which yields the most parsimonious model within one standard error of the minimum CV error.
(ii) Next, let $\{r_i\}_{i=1}^n$ be the Nuclear-LS residuals with c_λ selected via CV as in Step (i). As a rule-of-thumb, we set c_τ as the median absolute deviation of $\{r_i\}$, i.e. median$\{|r_i - \text{median}(r_i)|\}/0.6745$.
(iii) With c_τ determined after Step (ii), we choose the constant c_λ in the Nuclear-AH method again via five-fold CV under the one-standard-error rule.

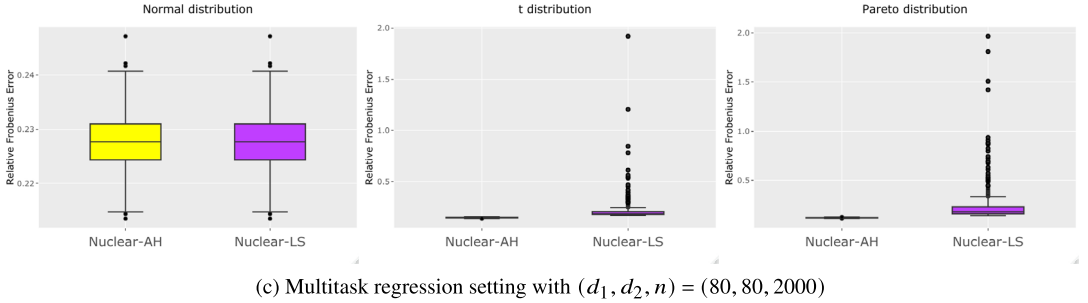
Under the matrix sensing and matrix completion setups, the data $\{(y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are generated from $y_i = \langle \boldsymbol{X}_i, \boldsymbol{\Theta}^*\rangle + \epsilon_i$, where ϵ_i follows one of the following three distributions: (i) $\mathcal{N}(0, 0.5^2)$—centered normal distribution with standard deviation 0.5 (lighted-tailed and symmetric), (ii) $t_{2.1}/8$—scaled t-distribution with 2.1 degrees of freedom (heavy-tailed and symmetric), and (iii) Par$(2,1)/8$—scaled Pareto distribution with scale parameter 1 and shape parameter 2 (heavy-tailed and asymmetric). For matrix sensing, we set $(d_1, d_2, n) = (50, 50, 1500)$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$ is such that $\text{rank}(\boldsymbol{\Theta}^*) = 5$ and all nonzero singular values of $\boldsymbol{\Theta}^*$ are 1, and the design matrix $\boldsymbol{X}_i$ consists of i.i.d. standard normal entries. For matrix completion, we set $(d_1, d_2, n) = (50, 50, 2000)$, $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$ is such that $\|\boldsymbol{\Theta}^*\|_{\mathrm{F}} = \sqrt{d_1 d_2}$ and $\text{rank}(\boldsymbol{\Theta}^*) = 5$, and $\boldsymbol{X}_i$ is uniformly sampled from $\{\boldsymbol{e}_j(d_1)\boldsymbol{e}_k^{\mathrm{T}}(d_2)\}_{1\le j\le d_1, 1\le k\le d_2}$. To implement LAMM, we use the initial estimates $\boldsymbol{\Theta}^{(0)} = \boldsymbol{0}$ and $\boldsymbol{\Theta}^{(0)} = (d_1 d_2/n)\sum_{i=1}^n y_i \boldsymbol{X}_i$, respectively, under the two setups. In the case of multitask regression, the data vectors $\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ are generated from $\boldsymbol{y}_i = (\boldsymbol{\Theta}^*)^{\mathrm{T}}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1\times d_2}$ is the same as in the matrix sensing setting, $\boldsymbol{x}_i \in \mathbb{R}^{d_1}$ are i.i.d. standard normal and



(a) Matrix sensing setting with $(d_1, d_2, n) = (50, 50, 1500)$



(b) Matrix completion setting with $(d_1, d_2, n) = (50, 50, 2000)$

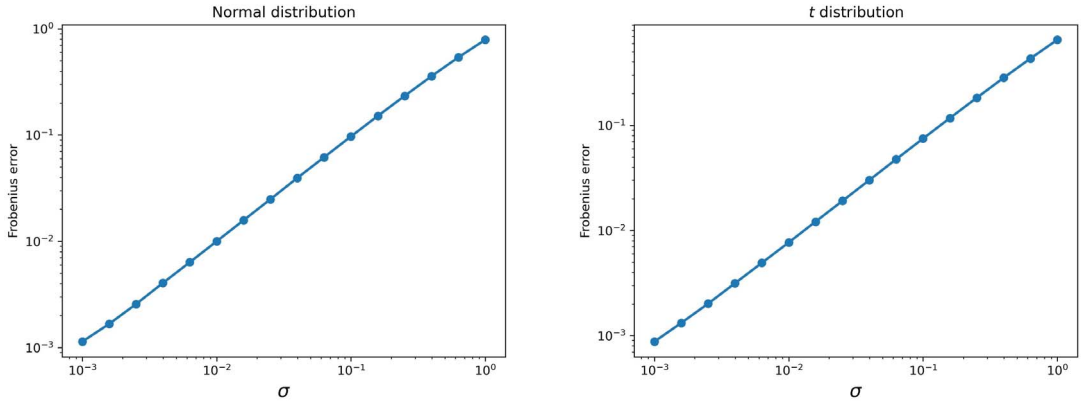


(c) Multitask regression setting with $(d_1, d_2, n) = (80, 80, 2000)$

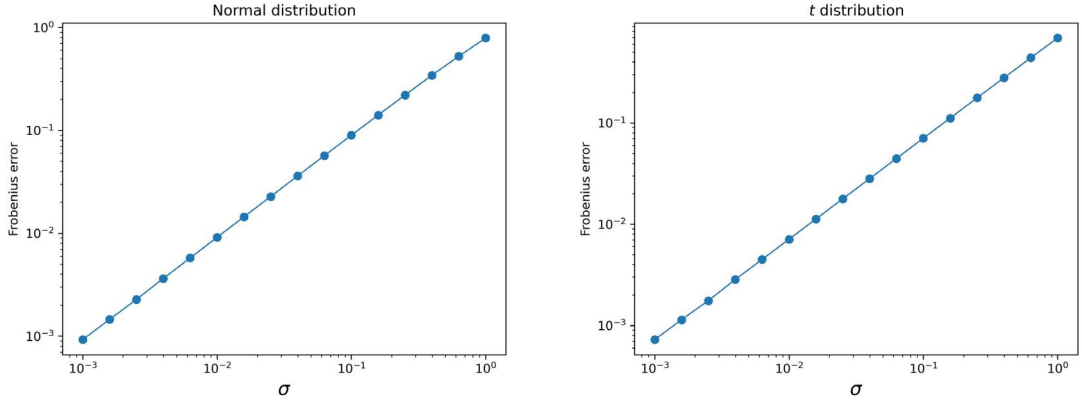
Figure 1. Boxplots of relative Frobenius errors $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}/\|\mathbf{\Theta}^*\|_{\mathrm{F}}$ (based on 500 repetitions) for the Nuclear-AH and Nuclear-LS estimators under the matrix sensing, matrix completion and multitask regression settings.

$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{id_2})^{\mathrm{T}}$ consists of i.i.d. entries following one of the above three errors distributions. In this case we set $d_1 = d_2 = 80$ and $n = 2000$.

Simulation results on the relative Frobenius error $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}/\|\mathbf{\Theta}^*\|_{\mathrm{F}}$, averaged over 500 repetitions, are presented in Table 1. To better demonstrate the robustness property of Nuclear-AH, Figure 1 shows the boxplots of (relative) Frobenius errors for the cross-validated Nuclear-LS and Nuclear-AH estimators under three error distributions. We see that Nuclear-LS and Nuclear-AH have almost identical performance when errors have symmetric and light tails, while the latter achieves considerably better performance under all three settings in the presence of heavy-tailed and/or asymmetric errors.



(a) Matrix sensing setting with $(d_1, d_2, n) = (50, 50, 2000)$



(b) Multitask regression setting with $(d_1, d_2, n) = (50, 50, 2000)$

Figure 2. Plots of Frobenius error $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}$ versus noise scale based on 200 simulations under the matrix sensing and multitask regression settings.

4.2. Convergence rate versus noise scale

In this section, we numerically examine the dependence of $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}$ on the noise scale under the matrix sensing and multitask regression settings. Our theoretical results, Theorem 3.3 and Theorem 3.5, indicate that in the exact low-rank case, $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}$ should be proportional to the noise scale σ, where $\sigma^2 = \mathbb{E}(\epsilon_i^2)$ or $\sigma^2 = \lambda_{\max}(\mathbb{E}(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^{\mathrm{T}}))$. To verify this, given a sequence of σ values ranging from 10^{-3} to 1, we generate ϵ_i from either $\sigma \cdot \mathcal{N}(0,1)$ or $\sigma \cdot t_{2.1}/16$. The specifications of $\boldsymbol{\Theta}^*$ and $\boldsymbol{X}_i$ or $\boldsymbol{x}_i$ are the same as in Section 4.1.

Under the matrix sensing setting, we set $(d_1, d_2, n) = (50, 50, 2000)$ and choose $\tau = 2\sigma\sqrt{n/d}$ and $\lambda = \sigma\sqrt{d/n}$ with $d = d_1 + d_2$. For multitask regression, we set $(d_1, d_2, n) = (100, 100, 3000)$ and choose $\tau = \sigma\sqrt{n/\log d}$ and $\lambda = 0.5\sigma\sqrt{d\log(d)/n}$. Figure 2 shows the plots of the Frobenius error versus noise scale, based on 200 replications, under these two settings and two error distributions. Consistent with the predictions of Theorems 3.3 and 3.5, we observe a nearly perfect linear growth in all four plots.

Acknowledgments

We thank the Editor, an Associate Editor, and two anonymous reviewers for their constructive comments and valuable suggestions, which have significantly helped us improve the quality of this work.

Funding

MY and WZ are supported in part by the NSF Grant DMS-2113409. QS is partially supported by Natural Sciences and Engineering Research Council of Canada (Grant RGPIN-2018-06484), a New Frontiers in Research Fund NFRFE-2019-00603, and a Data Sciences Institute Catalyst Grant.

Supplementary Material

Supplement to "Low-rank matrix recovery under heavy-tailed errors" (DOI: 10.3150/23-BEJ1675SUPP; .pdf). The supplementary material [42] contains the proofs of the theoretical results established in Section 3.

References

[1] Argyriou, A., Evgeniou, T. and Pontil, M. (2008). Convex multi-task feature learning. *Mach. Learn.* **73** 243–272.

[2] Burer, S. and Monteiro, R.D.C. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.* **95** 329–357. MR1976484 https://doi.org/10.1007/s10107-002-0352-8

[3] Cai, J.-F., Candès, E.J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20** 1956–1982. MR2600248 https://doi.org/10.1137/080738970

[4] Candès, E.J. and Plan, Y. (2009). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

[5] Candès, E.J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory* **57** 2342–2359. MR2809094 https://doi.org/10.1109/TIT.2011.2111771

[6] Candès, E.J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

[7] Chen, J., Liu, D. and Li, X. (2020). Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Trans. Inf. Theory* **66** 5806–5841. MR4158648 https://doi.org/10.1109/TIT.2020.2992234

[8] Chen, Y., Jalali, A., Sanghavi, S. and Caramanis, C. (2013). Low-rank matrix recovery from errors and erasures. *IEEE Trans. Inf. Theory* **59** 4324–4337.

[9] Dalalyan, A. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems* **32** 13188–13198.

[10] Elsener, A. and van de Geer, S. (2018). Robust low-rank matrix estimation. *Ann. Statist.* **46** 3481–3509. MR3852659 https://doi.org/10.1214/17-AOS1666

[11] Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972 https://doi.org/10.1111/rssb.12166

[12] Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841. MR3782385 https://doi.org/10.1214/17-AOS1568

[13] Fan, J., Wang, W. and Zhu, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Ann. Statist.* **49** 1239–1266. MR4298863 https://doi.org/10.1214/20-aos1980

[14] Goldberg, D., Nichols, D., Oki, B.M. and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM* **35** 61–70.
[15] Gross, D., Liu, Y.-K., Flammia, S.T., Becker, S. and Eisert, J. (2010). Quantum state tomography via compressed sensing. *Phys. Rev. Lett.* **105** 150401. https://doi.org/10.1103/PhysRevLett.105.150401
[16] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373
[17] Izenman, A.J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** 248–264. MR0373179 https://doi.org/10.1016/0047-259X(75)90042-1
[18] Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583 https://doi.org/10.3150/12-BEJ486
[19] Klopp, O., Lounici, K. and Tsybakov, A.B. (2017). Robust matrix completion. *Probab. Theory Related Fields* **169** 523–564. MR3704775 https://doi.org/10.1007/s00440-016-0736-y
[20] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894
[21] Li, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. *Constr. Approx.* **37** 73–99. MR3010211 https://doi.org/10.1007/s00365-012-9176-9
[22] Li, Y., Ma, T. and Zhang, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on Learning Theory* 2–47.
[23] Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A.B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896
[24] Luan, X., Fang, B., Liu, L., Yang, W. and Qian, J. (2014). Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion. *Pattern Recognit.* **47** 495–508.
[25] Ma, C., Wang, K., Chi, Y. and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.* **20** 451–632. MR4099988 https://doi.org/10.1007/s10208-019-09429-9
[26] Ma, J. and Fattahi, S. (2023). Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *J. Mach. Learn. Res.* **24** Paper No. [96], 84. MR4582518
[27] Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642
[28] Negahban, S. and Wainwright, M.J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348 https://doi.org/10.1214/10-AOS850
[29] Negahban, S. and Wainwright, M.J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649
[30] Recht, B., Fazel, M. and Parrilo, P.A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543 https://doi.org/10.1137/070697835
[31] Rohde, A. and Tsybakov, A.B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342 https://doi.org/10.1214/10-AOS860
[32] She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika* **104** 633–647. MR3694587 https://doi.org/10.1093/biomet/asx032
[33] Shen, Y., Li, J., Cai, J. and Xia, D. (2022). Computationally efficient and statistically optimal robust low-rank matrix estimation. arXiv:2203.00953.
[34] Sun, Q., Zhou, W.-X. and Fan, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461 https://doi.org/10.1080/01621459.2018.1543124
[35] Tan, K.M., Sun, Q. and Witten, D. (2022). Sparse reduced rank Huber regression in high dimensions. *J. Amer. Statist. Assoc.* To appear.
[36] Thompson, P. (2020). Outlier-robust sparse/low-rank least-squares regression and robust matrix completion. arXiv:2012.06750.
[37] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[38] Tong, T., Ma, C. and Chi, Y. (2021). Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.* **22** Paper No. 150, 63. MR4318506 https://doi.org/10.1080/15502287.2020.1856971

[39] Trefethen, L.N. and Bau, D. III (1997). *Numerical Linear Algebra*. Philadelphia, PA: SIAM. MR1444820 https://doi.org/10.1137/1.9780898719574

[40] Wang, B. and Fan, J. (2022). Robust matrix completion with heavy-tailed noise. arXiv:2206.04276.

[41] Wei, K., Cai, J.-F., Chan, T.F. and Leung, S. (2016). Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM J. Matrix Anal. Appl.* **37** 1198–1222. MR3543156 https://doi.org/10.1137/15M1050525

[42] Yu, M., Sun, Q. and Zhou, W.-X. (2024). Supplement to "Low-rank matrix recovery under heavy-tailed errors." https://doi.org/10.3150/23-BEJ1675SUPP

[43] Zhang, J., Fattahi, S. and Zhang, R. (2021). Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. In *Advances in Neural Information Processing Systems* **34** 5985–5996.

Received November 2022 and revised September 2023