Journal of Complexity 85 (2024) 101871

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Adaptive Huber trace regression with low-rank matrix parameter via nonconvex regularization ☆

Xiangyong Tan [a,b,1], Ling Peng [a,b,1], Heng Lian [c], Xiaohui Liu [b,a,*]

[a] School of Statistics and Data Science, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China
[b] Key Laboratory of Data Science in Finance and Economics, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China
[c] Department of Mathematics, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 10 December 2023
Received in revised form 4 May 2024
Accepted 30 May 2024
Available online 11 June 2024

MSC:
62J05
62H12

Keywords:
Huber trace regression model
Low-rank
Oracle inequality
Nonconvex regularization

ABSTRACT

In this paper, we consider the adaptive Huber trace regression model with matrix covariates. A non-convex penalty function is employed to account for the low-rank structure of the unknown parameter. Under some mild conditions, we establish an upper bound for the statistical rate of convergence of the regularized matrix estimator. Theoretically, we can deal with heavy-tailed distributions with bounded $(1+\delta)$-th moment for any $\delta > 0$. Furthermore, we derive the effect of the adaptive parameter on the final estimator. Some simulations, as well as a real data example, are designed to show the finite sample performance of the proposed method.

☆ Communicated by P. Przybyłowicz.

* Corresponding author at: Key Laboratory of Data Science in Finance and Economics, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China.

E-mail address: liuxiaohui@jxufe.edu.cn (X. Liu).

[1] The first two authors contributed equally.

https://doi.org/10.1016/j.jco.2024.101871

1. Introduction

Heavy-tailed data are often present in fields like finance, economics, environmental data analysis, etc., especially when high-frequency data are involved. It has been well-known that the conventional least squares (LS) method is sensitive to the existence of outliers when model errors are possibly not Gaussian but heavy-tailed. In the case of linear models, regression estimators based on the least-squares loss are theoretically and empirically suboptimal when non-Gaussian errors are present. A deviation analysis conducted by [2] demonstrates that the deviation of the empirical mean can be significantly worse for non-Gaussian samples compared to Gaussian ones.

To account for the presence of heavy-tailed errors, a reliable and widely used method is the least absolute deviation (LAD) regression. Introduced by Roger Joseph Boscovich in 1757 [5,18], LAD regression can be viewed as a special case of quantile regression [18]. By employing an absolute loss function, LAD regression places relatively less emphasis on errors with larger absolute values, unlike least squares (LS) regression, which uses a squared loss function. This makes LAD regression more resistant to outliers in the response variable. Notably, LS regression may yield efficient results when model errors are potentially Gaussian. In contrast, LAD regression is robust against potential outliers when the underlying distribution of model errors is heavy-tailed. Huber [16] proposed a piecewise loss function, known as the Huber loss, which combines the loss functions used in LS and LAD regressions. The Huber regression stands out for its ability to balance efficiency and robustness under both Gaussian and heavy-tailed errors.

Since its introduction, the Huber loss has become a significant robust criterion for estimating parameters. The asymptotic properties of Huber estimators have been extensively studied in fixed or low-dimensional settings [17,38]. More recently, [7], [33], and [30] have made novel findings on adaptive robust estimation based on the Huber loss for high dimensional mean regression. Specifically, in the presence of asymmetric errors, [7] and [30] investigate Huber-type estimators and provide non-asymptotic estimation bounds. Under symmetry around zero of the error assumption, [22] studied the inference for high-dimensional linear models without an intercept term under the weighted Huber loss. In contrast, [13] introduced a robust post-selection inference approach for the regression coefficients in a high-dimensional linear model with an intercept term, employing the Huber loss, particularly in cases where the error distribution is heavy-tailed and asymmetric.

Much of the above literature on the Huber regression focuses on the case with vector-type covariates. However, in practice, the range of available data has expanded beyond numerical data, such as panel data, two-dimensional digital imaging, and electroencephalography. Some of these data are presented in matrix form, rather than as a vector. A distinctive characteristic of matrix-type data is their inherent structure and the ability to capture the correlation between rows and columns simultaneously. Consequently, analyzing matrix-type data through simple vectorization may prove to be inefficient.

When matrix-type covariates are involved, the simplest but most useful tool for analysis is the following trace regression model:

$$Y_i = \langle X_i, \Theta \rangle + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where $\langle X_i, \Theta \rangle = \mathrm{tr}(\Theta^\top X)$ denotes the trace operator, $X \in \mathbb{R}^{d_1 \times d_2}$ is a matrix of explanatory variables with dimensions $d_1 \geq 1$ and $d_2 \geq 1$, $\Theta \in \mathbb{R}^{d_1 \times d_2}$ denotes the matrix of unknown regression coefficients, $Y \in \mathbb{R}$ is the response and $\varepsilon \in \mathbb{R}$ is the model error with zero mean. Note that $\mathrm{tr}(\Theta^\top X) = vec(\Theta)^\top vec(X)$, where $vec(\cdot)$ denotes the vectorized version of a given matrix sense. In this sense, the model (1) can be considered a direct extension of the well-studied linear regression model. To make the estimation feasible, certain types of sparsity must be assumed, as the dimension of $vec(\Theta)$, which is equal to $d_1 d_2$, may be considerably high even when d_1 and/or d_2 are/is relatively small.

To estimate the unknown parameter Θ in the model (1), the main method involves assuming a certain level of sparsity in the rank of the parameter matrix (Θ) and subsequently implementing the nuclear norm penalty to reduce the magnitude of the estimator. This approach is rooted in the understanding that to determine a rank-r matrix $\Theta \in R^{d_1 \times d_2}$, only r left and right singular vectors, along with r singular values, are necessary. These singular vectors and values correspond to $r(d_1 + d_2 - 1)$ degrees of freedom, without taking orthogonality into account [9]. Therefore, low-rank matrices typically possess significantly fewer degrees of freedom than their ambient dimensions of $d_1 d_2$. Based on the low-rank assumption, [19] obtained a general sharp oracle inequality for the nuclear norm penalized estimator of the trace regression model. Fan et al. [9] further studied the generalized trace regression with a near-low-rank regression coefficient matrix.

In real data analysis, data often exhibit both low-rank and sparse structures. Researchers have introduced additional sparsity assumptions to capture these structures on the matrix Θ, which is simultaneously low-rank and element-wise sparse. For instance, [26] and [3] studied the asymptotic properties of the estimate in mean trace regression by considering a composite penalty that combines the nuclear norm and the L_1 norm. Peng et al. [29] further studied the linear trace regression with β-mixing errors. However, if the covariate X has the property that variables in the same row (or column) share similar information or are associated with a common factor, sparse elements may not be appropriate. In light of this, [37] investigated the oracle inequality in the trace regression model with simultaneous low rank and row (or column) sparsity using the nuclear norm and group lasso penalties. Tan et al. [31] extended this work to a quantile linear model and derived an upper bound for the convergence rate.

In model (1), most previous studies have utilized a nuclear norm penalty to obtain a low-rank estimator. Although the computational attractiveness of the convexity of the nuclear norm penalty is evident, there is still a bias present in its estimator. As a result, scholars have proposed the use of non-convex penalties, such as the smoothly clipped absolute deviation penalty (SCAD, [6]), mini-max concave penalty (MCP, [35]), and capped L_1 penalty [36]. Extensive research has demonstrated that, compared to a convex relaxation with the L_1 norm, employing an appropriate non-convex penalty method enables achieving sparse estimation with fewer measurements and greater robustness against noise [4]. However, there has been limited research conducted on the application of non-convex penalty in trace regression under Huber loss.

This work presents a novel procedure that employs a non-convex penalty in conjunction with Huber loss in a linear trace regression model. The proposed work shows novelty in several aspects: (1) Owning to the utilization of Huber loss, our approach is capable of effectively handling heavy-tailed or asymmetric errors, which only have finite $(1 + \delta)$-th moments. (2) By incorporating non-convex nuclear norm regularization to address the low-rank structure, we establish the convergence rate for the coefficient matrix. (3) As the nonconvex and nonsmooth characteristics of the objective function, we extend the local adaptive majorize-minimization algorithm developed in [8] to estimate the unknown parameters. In this algorithm, we employ a data-driven approach to determine the robustification parameter, which aims to balance robustness and bias. This study is practically motivated by real-world applications. The Beijing Air Quality dataset, as described in Section 4, consists of a 24×21 matrix serving as the predictor variable, coupled with the response variable representing the daily aggregated count of PM2.5, which is characterized by its heavy and slightly asymmetric distribution. We aim to investigate the relationship between PM2.5 and these matrix covariates. Notably, while a matrix predictor can be transformed into a vector format, such manipulation may compromise the inherent structure and result in the loss of valuable information. Motivated by this example, this paper focuses on the regularized Huber matrix regression, where a matrix is employed as the predictor, complemented by a scalar response variable.

The remainder of the paper is organized as follows. In Section 2, we propose a non-convex penalized estimator utilizing Huber loss and explicitly derive the statistical rate of this estimator. Section 3 presents the algorithm used and provides the finite-sample simulation results. Section 4 presents the results of the real data analysis. The fifth section concludes this paper. Finally, we give detailed proofs of the theorems and the associated technical details.

2. Methodology and main results

2.1. Notations

For convenience, we first introduce some useful notations that are commonly utilized in the existing literature related to trace regression. For a random vector $\boldsymbol{x} \in \mathbb{R}^d$ $(d \geq 1)$, we define the sub-Gaussian norm and subexponential norm as $\|\boldsymbol{x}\|_{\Psi_2} = \sup_{\boldsymbol{v} \in \mathcal{S}^{d-1}} \left\|\boldsymbol{v}^T \boldsymbol{x}\right\|_{\Psi_2}$ and $\|\boldsymbol{x}\|_{\Psi_1} = \sup_{\boldsymbol{v} \in \mathcal{S}^{d-1}} \left\|\boldsymbol{v}^T \boldsymbol{x}\right\|_{\Psi_1}$, respectively, where $\mathcal{S}^{d-1}$ denotes the unit ball in $\mathbb{R}^d$. For matrices $A_1, A_2 \in \mathbb{R}^{d_1 \times d_2}$, denote their Frobenius inner product as $\langle A_1, A_2 \rangle = \operatorname{tr}(A_1^\top A_2)$. For any $A \in \mathbb{R}^{d_1 \times d_2}$, denote $\{\sigma_k(A)\}_{k=1}^d$ as the sequence of nondecreasing singular values, where $d = \min\{d_1, d_2\}$. Denote $vec(A) \in \mathbb{R}^{d_1 d_2}$ as the vector of all the elements from A column by column. Denote $\|A\|_{2,1} = \sum_{j=1}^{d_2} \|A_{\cdot j}\|$ and $\|A^\top\|_{2,1} = \sum_{i=1}^{d_1} \|A_{i\cdot}\|$ with $A_{i\cdot}$ and $A_{\cdot j}$ being the i-th row and the j-th column, respectively. Furthermore, we define the operator norm (spectral norm) $\|A\|_{op}$, the nuclear norm (trace norm) $\|A\|_*$, and the Frobenius norm $\|A\|_F$ of A as $\|A\|_{op} = \max_{x \in \mathbb{R}^{d_2}} \frac{\|Ax\|}{\|x\|}$, $\|A\|_N = \sum_{k=1}^d \sigma_k(A)$, and $\|A\|_F = \sqrt{\langle A, A \rangle} = \left(\sum_{k=1}^d \sigma_k(A)^2\right)^{1/2}$, respectively. For any subspace $\mathcal{V} \subset \mathbb{R}^{d_1 \times d_2}$, define its orthogonal space as $\mathcal{V}^\perp = \{Q : \forall S \in \mathcal{V}, \langle Q, S \rangle = 0\}$. Furthermore, for $\{a_n\}, \{b_n\}$ with $a_n, b_n > 0$, by $a_n \asymp b_n$ we mean that a_n / b_n is bounded away from both zero and infinity as $n \to \infty$.

2.2. Methodology and main results

Suppose that random observations $\{(X_i, Y_i)\}_{i=1}^n$ are generated from (1), it is important to acknowledge the possibility of infinite variance for ε_i. Consequently, the use of the l_2 loss may not be appropriate. Furthermore, employing Huber regression with a fixed tuning constant may result in significant estimation bias. To address this limitation, we suggest utilizing the Huber loss with an adaptive robustification parameter, simultaneously enabling the attainment of robustness and (asymptotic) unbiasedness. By [16], we propose to use the Huber loss, i.e.,

$$l_\alpha(x) = \begin{cases} x^2, & |x| \leq \alpha, \\ 2\alpha|x| - \alpha^2, & |x| > \alpha, \end{cases} \tag{2}$$

where $\alpha > 0$ is referred to as the robustification parameter, which plays a crucial role in balancing the trade-off between bias and robustness. By Huber loss, we have the following empirical loss function:

$$L_{n,\alpha}(\Theta) = \frac{1}{n} \sum_{i=1}^n l_\alpha \left(Y_i - \langle X_i, \Theta \rangle\right). \tag{3}$$

To address the estimation challenge in the high-dimensional scenario, it is generally assumed that the matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ possesses a low rank, where $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ denotes the true value of Θ. In this paper, we assume that Θ^* is exactly low-rank. To explore the low-rank structure of Θ^*, we employ the following penalized loss function,

$$\hat{\Theta} := \underset{\Theta \in \mathbb{R}^{d_1 \times d_2}}{\arg\min} \left\{L_{n,\alpha}(\Theta) + \mathcal{R}_\lambda(\Theta)\right\} \tag{4}$$

to estimate the unknown parameter Θ^*. $\mathcal{R}_\lambda : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ is a regularizer depending on the regularization parameter λ, written as

$$\mathcal{R}_\lambda(\Theta) = \sum_{i=1}^d p_\lambda(\sigma_i(\Theta)),$$

with $d = \min(d_1, d_2)$ and $\sigma_1(\Theta) \geq \cdots \geq \sigma_d(\Theta) \geq 0$ being singular values of Θ in descending order. The regularizer $p_\lambda(\cdot)$, which imposes some sparsity constraint on the estimator, can be non-convex.

Several nonconvex regularizers have been suggested in the literature, including SCAD [6] and MCP [35]. When applying the SCAD and MCP regularizers, [12] suggested that it can be decomposed into

$$\mathcal{R}_\lambda(\Theta) = \lambda\|\Theta\|_* + \mathcal{Q}_\lambda(\Theta), \tag{5}$$

where $\mathcal{Q}_\lambda(\Theta) = \sum_{i=1}^{d} q_\lambda(\sigma_i(\Theta))$. This decomposition will play an important role in the proof of the main results of this paper. Under the decomposition of the regularizer in (5), the estimator in (4) can be rewritten as

$$\hat{\Theta} := \underset{\Theta\in\mathbb{R}^{d_1\times d_2}}{\arg\min} \left\{\tilde{L}_{n,\alpha,\lambda}(\Theta) + \lambda\|\Theta\|_*\right\}, \tag{6}$$

where $\tilde{L}_{n,\alpha,\lambda}(\Theta) = L_{n,\alpha}(\Theta) + \mathcal{Q}_\lambda(\Theta)$.

In this paper, we are interested in studying the statistical rate of $\|\hat{\Theta} - \Theta^*\|_F$. Let

$$\Theta^*_\alpha = \underset{\Theta\in\mathbb{R}^{d_1\times d_2}}{\arg\min} \left\{\mathbb{E} l_\alpha(Y - \langle X, \Theta\rangle)\right\}.$$

In general, Θ^*_α differs from Θ^*. According to [7], the statistical error can be decomposed into the approximation error $\Theta^*_\alpha - \Theta^*$ and the estimation error $\hat{\Theta} - \Theta^*_\alpha$. The statistical rate of $\|\hat{\Theta} - \Theta^*\|_F$ is then bounded by

$$\|\hat{\Theta} - \Theta^*\|_F \le \|\hat{\Theta} - \Theta^*_\alpha\|_F + \|\Theta^* - \Theta^*_\alpha\|_F.$$

To derive the convergence rate of $\|\hat{\Theta} - \Theta^*\|_F$, we need to specify the following regularity conditions.

(C1) Regression errors ε_i satisfy $\mathbb{E}[\varepsilon_i|X_i] = 0$ and $\mathbb{E}[|\varepsilon_i|^{1+\delta}|X_i] < \infty$ almost surely for some $\delta > 0$.
(C2) $0 < \rho_l \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le \rho_u < \infty$, where $\Sigma = \mathbb{E}(vec(X)(vec(X))^\top)$. $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the minimum and maximum eigenvalues of Σ, respectively.
(C3) $vec(X)$ is a sub-Gaussian random vector with $K_X = \max_{v\in\mathcal{S}^{d_1\times d_2-1}} \|(vec(X)^\top v\|_{\psi_2} < \infty$.

Conditions (C1)-(C3) are some regular conditions. The following Proposition gives the upper bound of the approximation bias, which is of order $\alpha^{-\delta}$.

Proposition 1. *Under Assumptions (C1)-(C3), the approximation error satisfies*

$$\|\Theta^*_\alpha - \Theta^*\|_F \le C\rho_l^{-1}\rho_u^{1/2}(K_\varepsilon^{1/2} + K_X^{1+\delta})\alpha^{-\delta},$$

where $C > 0$ is some constant.

Next, we establish the main results of the convergence rate of our proposed estimator. Furthermore, we impose several regularity conditions on the non-convex penalty $\mathcal{R}_\lambda(\cdot)$, in terms of functions $p_\lambda(\cdot)$ and $q_\lambda(\cdot)$.

(C4) For the non-convex penalty $\mathcal{R}_\lambda(\cdot)$, we have following assumptions:

(i). The function $p_\lambda(t)$ is non-decreasing and differentiable for $t \ne 0$ and sub-differentiable at $t = 0$ with $\lim_{t\to 0^+} p'_\lambda(t) = \lambda$;
(ii). There exists a positive constant $\gamma > 0$ such that $p'_\lambda(t) = 0$, for all $t \ge \gamma\lambda$;
(iii). q_λ is concave with $q_\lambda(0) = q'_\lambda(0) = 0$. For $t' \ge t \ge 0$, there exists a constant $\eta_- \ge 0$ such that $q'_\lambda(s) - q'_\lambda(t) \ge -\eta_-(s - t)$;
(iv). For $t > 0$, $|q'_\lambda(t)| \le \lambda$.

Remark 1. The conditions in Assumption (C4) are similar to those proposed in [10,24], which are satisfied by many widely used non-convex penalties such as SCAD and MCP. Note that the last condition (iv) is the same as condition (vi) proposed in [21], which is a generalization of the weak convexity assumption [23].

For the unknown parameter matrix Θ^*, the singular value decomposition (SVD) can be expressed as follows: $\Theta^* = U_r \Gamma^* V_r^\top$, where Γ^* is a diagonal matrix in $\mathbb{R}^{r\times r}$ containing the singular values $\sigma_1(\Theta^*), \cdots, \sigma_r(\Theta^*)$ in nonincreasing order. The matrices $U_r = (u_1, \ldots, u_r) \in \mathbb{R}^{d_1\times r}$ and $V_r = (v_1, \ldots, v_r) \in \mathbb{R}^{d_2\times r}$ represent the columns of the left and right singular vectors, respectively. According to U_r, V_r, we define the following two subspaces of $\mathbb{R}^{d_1\times d_2}$:

$$\mathcal{M} := \left\{ \Theta \in \mathbb{R}^{d_1\times d_2} : \mathrm{row}(\Theta) \subseteq V_r\ ,\ \mathrm{col}(\Theta) \subseteq U_r \right\},$$

$$\overline{\mathcal{M}}^\perp := \left\{ \Theta \in \mathbb{R}^{d_1\times d_2} : \mathrm{row}(\Theta) \perp V_r\ ,\ \mathrm{col}(\Theta) \perp U_r \right\},$$

where row$(\cdot)$ and col$(\cdot)$ denote row space and the column space, respectively. On the other hand, the second equation above defines the subspace $\overline{\mathcal{M}}$ implicitly via taking the orthogonal complement. For any matrix $B \in \mathbb{R}^{d_1\times d_2}$, we define the projector onto the linear space spanned by the first r columns of the left (right) singular vectors as

$$\mathcal{P}_{\mathcal{M}}(B) = U_r U_r^\top B V_r V_r^\top,$$

and the projector onto the orthogonal space is given by

$$\mathcal{P}_{\overline{\mathcal{M}}^\perp}(B) = (I_{d_1} - U_r U_r^\top) B (I_{d_2} - V_r V_r^\top).$$

To begin with, we impose the restricted strong convexity (RSC) conditions on the empirical loss function over a restricted set. This assumption assumes that the remainder of the first-order Taylor expansion of $L_{n,\alpha}(\Theta)$ has a quadratic lower bound. We define the following subset, which is a cone with a restricted set of directions,

$$\mathcal{C} = \left\{ \Delta \in \mathbb{R}^{d_1\times d_2} \,|\, \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\Delta)\|_* \le 5\|\mathcal{P}_{\overline{\mathcal{M}}}(\Delta)\|_* \right\}.$$

Definition 1. (Restricted Strong Convexity (RSC)) Given the constrained set $\mathcal{C}$ defined above, there exist positive constants κ_l such that

$$L_{n,\alpha}(\Theta + \Delta) - L_{n,\alpha}(\Theta) - \langle \nabla L_{n,\alpha}(\Theta), \Delta \rangle \ge \kappa_l \|\Delta\|_F^2. \tag{7}$$

Here, $\nabla L_{n,\alpha}(\Theta)$ is defined as the gradient of loss function $L_{n,\alpha}(\Theta)$ with respect to Θ, i.e.

$$\nabla L_{n,\alpha}(\Theta) = -\frac{1}{n}\sum_{i=1}^{n} l'_\alpha(Y_i - \langle X_i, \Theta\rangle) X_i,$$

where $l'_\alpha(x) = 2\mathrm{sign}(x)\min\{|x|, \alpha\}$ for all $x \in \mathbb{R}$ with sign $(\cdot)$ being the sign function. The RSC condition has been thoroughly discussed in previous literature [23,27]. This condition guarantees the strong convexity of the loss function within a restricted set $\mathcal{C}$ and helps to control the estimation error $\|\hat{\Theta} - \Theta^*\|_F$.

Theorem 1 *(Deterministic Bound). Assume that the conditions (**C1**)-(**C4**) and the RSC condition (7) hold. If* $\hat{\Theta} - \Theta^* \in \mathcal{C}$ *and* $\lambda \ge 2\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}$, *the estimator in (4) achieves the estimation error*

$$\|\hat{\Theta} - \Theta^*\|_F \le \frac{5/2\lambda\sqrt{2r}}{2\kappa_l - \eta_-}. \tag{8}$$

Theorem 1 is a deterministic result that relies on the RSC condition. In the subsequent analysis, we will present a proposition that demonstrates that the RSC condition is satisfied under some conditions. Write

$$\nu_{i,\delta} = \mathbb{E}[|\varepsilon_i|^{1+\delta}|X_i] \qquad \text{and} \qquad \nu_\delta = \frac{1}{n}\sum_{i=1}^{n} \nu_{i,\delta}.$$

Assuming $\nu_\delta < \infty$ for some $\delta > 0$. Furthermore, we let

$$\mathcal{B}^*(R) := \left\{\Theta \in R^{d_1 \times d_2} : \|\Sigma^{1/2} vec(\Theta - \Theta^*)\| \le R\right\}.$$

The following propositions show that the adaptive Huber loss function $L_{n,\alpha}$ satisfies the local RSC condition over $\mathcal{B}^*(R) \cap \mathcal{C}$ with high probability.

Proposition 2. *Suppose that the conditions (**C1**)-(**C3**) hold. Assume that $\mathbb{E}\langle X, \Theta\rangle^4 \le C\left(\mathbb{E}\langle X, \Theta\rangle^2\right)^2$ for some constant $C > 0$ and all $\Theta \in \mathbb{R}^{d_1 \times d_2}$. Let $\alpha \ge 4\max\{2C^{1/4}R, \nu_\delta^{1/(1+\delta)}\}$, and let $n \gtrsim \rho_l^{-1} r(\alpha/R)^2(d_1 + d_2)$. Then, with probability at least $1 - e^{-(d_1+d_2)}$,*

$$L_{n,\alpha}(\Theta + \Delta) - L_{n,\alpha}(\Theta) - \langle \nabla L_{n,\alpha}(\Theta), \Delta\rangle \ge \frac{\rho_l}{4}\|\Theta - \Theta^*\|_F^2, \tag{9}$$

uniformly over $\Theta \in \mathcal{B}^*(R) \cap \mathcal{C}$.

To obtain asymptotic results on the estimator $\hat{\Theta}$, we need to investigate the convergence rate of $\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}$.

Proposition 3. *Under assumption (**C1**)-(**C3**), for some constant $C > 0$, we have*

$$\|\nabla L_{n,\alpha}(\Theta^*)\|_{op} \le C\left(\sqrt{\frac{\nu_\delta \alpha^{1-\delta}(d_1 + d_2)}{n}} + \frac{\alpha(d_1 + d_2)}{n} + \nu_\delta \alpha^{-\delta}\right), \tag{10}$$

with probability at least $1 - 2 \times 7^{-(d_1+d_2)}$.

We are ready to present the main result on the adaptive Huber trace estimator in high dimensions with the above results.

Theorem 2. *Assume that the conditions (**C1**)–(**C4**) and the RSC condition (7) hold. Suppose that $\hat{\Theta} - \Theta^* \in \mathcal{C}$, then for the robustification parameter α and the regularization parameter λ that satisfy*

$$\alpha \asymp \left(\frac{n}{d_1 + d_2}\right)^{\max\{1/(1+\delta), 1/2\}} \qquad \text{and} \qquad \lambda \asymp \left(\frac{d_1 + d_2}{n}\right)^{\min\{\delta/(1+\delta), 1/2\}},$$

the estimator $\hat{\Theta}$ in (4) achieves estimation errors

$$\|\hat{\Theta} - \Theta^*\|_F \le C\frac{\sqrt{r}}{2\kappa_l - \eta_-}\left(\frac{d_1 + d_2}{n}\right)^{\min\{\delta/(1+\delta), 1/2\}} \tag{11}$$

with probability at least $1 - 2 \cdot 7^{-(d_1+d_2)}$, where C is some absolute constant.

Remark 2. When the variance of ε is finite, i.e. $\delta \ge 1$, $\alpha \asymp \sqrt{n/(d_1 + d_2)}$ and $\lambda \asymp \sqrt{(d_1 + d_2)/n}$, the order of $\|\hat{\Theta} - \Theta^*\|_F$ is $\sqrt{r(d_1 + d_2)/n}$. However, when $0 < \delta < 1$, $\alpha \asymp (n/(d_1 + d_2))^{1/(1+\delta)}$ and $\lambda \asymp ((d_1 + d_2)/n)^{\delta/(1+\delta)}$. In this case, the order of $\|\hat{\Theta} - \Theta^*\|_F$ becomes $\sqrt{r}\,((d_1 + d_2)/n)^{\delta/(1+\delta)}$, which is slower than the order when $\delta \ge 1$.

3. Simulation study

This section proposes a computationally efficient algorithm for estimating matrix parameters and investigates its performance in finite sample scenarios using simulated experiments. First, we provide a detailed explanation of the algorithm's implementation. Subsequently, we present the simulation results to show its effectiveness. The algorithm is developed through a hybrid programming approach utilizing both R and C++ languages.

3.1. Computational algorithm

Being aware that the optimization of (6) is not trivial, we specify the details of the computational algorithm in obtaining the regularized estimates. The general structure of the computational algorithm is along the same line as that of [8], which developed a so-called local adaptive majorize-minimization (LAMM) algorithm for the adaptive Huber regression when vector-type covariates are involved. But slight differently, we locally majorize $L_{n,\alpha}(\Theta)$ in (3) by an isotropic quadratic function

$$g_k(\Theta|\Theta^{(k)}) = L_{n,\alpha}(\Theta^{(k)}) + \left\langle \nabla L_{n,\alpha}(\Theta^{(k)}), \Theta - \Theta^{(k)} \right\rangle + \frac{\phi_k}{2} \left\langle \Theta - \Theta^{(k)}, \Theta - \Theta^{(k)} \right\rangle, \tag{12}$$

where ϕ_k is a quadratic such that $g_k(\Theta^{(k+1)}|\Theta^{(k)}) \geq L_{n,\alpha}(\Theta^{(k+1)})$.

Similarly, we locally majorize $p_\lambda(\sigma_i(\Theta))$ by

$$p_\lambda(\sigma_i(\Theta^{(k)})) + \omega_i^k(\sigma_i(\Theta) - \sigma_i(\Theta^{(k)})), \tag{13}$$

where $\omega_i^k \in \partial p_\lambda(\sigma_i(\Theta^k))$. Then, by (12) and (13), we update $\Theta^{(k+1)}$ by solving

$$\Theta^{(k+1)} \arg\min_{\Theta} \left\{ \left\langle \nabla L_{n,\alpha}(\Theta^{(k)}), \Theta - \Theta^{(k)} \right\rangle + \frac{\phi_k}{2} \left\langle \Theta - \Theta^{(k)}, \Theta - \Theta^{(k)} \right\rangle + \sum_{i=1}^{d} \omega_i^k \sigma_i(\Theta) \right\}. \tag{14}$$

Let $\Theta = U_r \Sigma V_r^T$ and $S(\Theta, \omega) = U_r S_\omega(\Sigma) V_r^T$, where $S_\omega(\Sigma) = diag\{(\Sigma_{ii} - \omega_i)^+\}$ and $(\cdot)^+ = \max\{\cdot, 0\}$. It can be shown that (14) has a closed form:

$$\Theta^{(k+1)} = \mathcal{T}_{\phi_k}(\Theta^k) = S(\Theta^k - \phi_k^{-1} \nabla L_{n,\alpha}(\Theta^{(k)}), \phi_k^{-1} \omega^k), \tag{15}$$

where $\omega^k = (\omega_1^k, \cdots, \omega_d^k)$.

We formally summarize the computational algorithm in Algorithm 1 as follows:

Algorithm 1 LAMM algorithm for adaptive trace Huber regression.

1. Input: $\{X_i, Y_i\}_{i=1}^n$, λ.
2. Initialize: ϕ_0, $\Theta^{(0)}$, γ, ϵ.
3. for $k = 1, \cdots$ **until** $\|\Theta^{(k+1)} - \Theta^{(k)}\|_F \leq \epsilon$ **do**
4. Repeat
5. $\Theta^{(k+1)} \leftarrow \mathcal{T}_{\phi_k}(\Theta^k)$
6. If $g_k(\Theta^{(k+1)}|\Theta^{(k)}) < L_{n,\alpha}(\Theta^{(k+1)})$ **then** $\phi_k \leftarrow \gamma\phi_k$
7. Until $g_k(\Theta^{(k+1)}|\Theta^{(k)}) \geq L_{n,\alpha}(\Theta^{(k+1)})$
8. Return $\{\Theta^{(k+1)}, \phi_{k+1} = \max\{\phi_0, \gamma^{-1}\phi_k\}\}$
9. end for
10. Output: $\hat{\Theta} = \Theta^{(k+1)}$

It is noted that in Algorithm 1, we start from a small parameter $\phi_k = \phi_0$ and then successfully inflate ϕ_k by a factor $\gamma > 1$, say $\gamma_u = 1.1$. For the Huber loss parameter α, similar to [25], we update α at the beginning of each iteration in Algorithm 1. Let $\hat{R}^k = (\hat{r}_1^k, \cdots, \hat{r}_n^k)$, where $\hat{r}_i^k = Y_i - \langle X_i, \hat{\Theta}_{k-1} \rangle$, and $\hat{\Theta}_{k-1}$ is obtained from the $(k-1)$-th iteration of Algorithm 1. We define

$$mad(\hat{R}^k) = \{\Phi^{-1}(0, 75)\}^{-1} \text{median}(|\hat{R}^k - \text{median}(\hat{R}^k)|),$$

as the median absolute deviation of residuals. We start with setting

$$\alpha_0 = \{\Phi^{-1}(0,75)\}^{-1}\text{median}(|Y - \text{median}(Y)|).$$

At the k-th iteration of Algorithm 1, we update α by $\alpha_k = 0.125 \cdot mad(\hat{R}^k)\sqrt{n/\log(npq)}$.

3.2. Simulation results

We consider the random observations (X_i, Y_i), $i = 1, 2, ..., n$, generated from the model (1) with $X_i \sim N(0_{d\times d},\ I_d \otimes I_d)$ and $\Theta = D_1 D_2^\top$, where $D_1 \in R^{d\times K}$ and $D_2 \in R^{d\times K}$. In the sequel, we assume that $K = 2$, and the elements of D_1, D_2 are independently and identically generated from $N(3, 0.5)$. For the random model errors ϵ_i, we investigate the following four Scenarios:

Scenario 1. $\varepsilon_i = e_i - E(e_i)$, where $e_i \sim \log N(0, 4)$.
Scenario 2. $\varepsilon_i \sim t(1.5)$.
Scenario 3. $\varepsilon_i = e_i - E(e_i)$, where $e_i \sim Par(1, 1.6)$, and $Par(x_m, \alpha)$ means Pareto distribution with scale x_m and shape $\alpha > 0$.
Scenario 4. $\varepsilon_i \sim N(0, 1)$.

In Scenario 1, it should be noted that the distribution of model errors is not symmetrical, whereas in Scenarios 2 and 3, the model errors exhibit heavy-tailed characteristics. In all Scenarios, it is assumed that the error term ε_i is independent of the variable X_i.

It is noted that in the optimization of (6), the tuning number λ needs to be determined empirically by using some data-driven method. It is computationally time-consuming. Therefore, to relieve the computational burden, similar to [10] and [31], we use a validation set of size $100 \times n$ for tuning. The tuning parameter λ was selected for Huber loss and the L_2 loss by minimizing the validation error $\sum_{i\in validation}(Y_i - \hat{Y}_i)^2$. In the simulation, we set $d = 4, 8, 15$ and $n = 500, 800, 1000, 2000, 4000$. We report the quantities $D_{rk} = |\text{rank}(\hat{\Theta}) - K|$ and the median $ERR = \|\hat{\Theta} - \Theta^*\|_F$ based on 1000 simulations.

The results in $ERR = \|\Theta - \hat{\Theta}\|_F$ and D_{rk} for adaptive Huber trace regression and the trace least squares estimators, which average over 1000 simulations, are summarized in Tables 1-4. With heavy-tailed errors following log-normal, Student's t, and Pareto distributions, the adaptive Huber trace regression significantly outperforms the least squares. In the case of normal distribution noise, the adaptive Huber estimators perform as well as the least squares. These empirical results reveal that adaptive Huber trace regression prevails in various scenarios. In contrast to the LASSO, the simulation outcomes indicate that the SCAD or MCP penalty exhibits superior performance, particularly when the sample size (n) is small and the dimensionality (d) is large.

Based on the results presented in Fig. 1, it is evident that $\|\Theta - \hat{\Theta}\|_F$ decreases as the logarithm of the sample size n increases across various dimensions d. Furthermore, the slope of this decay is approximately -1/2, a finding that aligns with the expected order of n as indicated by the statistical rate derived for $\hat{\Theta}$.

4. Real data analysis

In this section, we use the proposed method to analyze Beijing air quality data, which is currently available on the website (https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data). The data set consists of continuously collected hourly measurements from 2013 to 2017. Following the methodology described in [34], any records with missing values are excluded, resulting in a dataset of 1035 complete records spanning 48 months. Each observation, denoted as $X_i \in R^{24\times 21}$, represents a 24-hour observation of various pollutants, including SO2, NO2, temperature (TEMP), pressure (PRES), dew point temperature (DEWP), wind speed (WSPM), and their second-order interactions. The response variable is the aggregated daily count of PM2.5. We incorporated XGBoost and RandomForest (RF) into our study for comparative analysis. The process involved vectorizing the predictors for XGBoost and Random Forest models, followed by training using the training data.

Table 1
Simulation results for Scenario 1.

n	Methods	$d=4$		$d=8$		$d=15$	
		ERR	D_{rk}	ERR	D_{rk}	ERR	D_{rk}
500	Huber_Lasso	0.2363(0.0674)	0.006	0.4157(0.0982)	0.001	0.9407(0.1890)	0.001
	Huber_Scad	0.2132(0.0527)	0.006	0.3530(0.0701)	0.001	0.5717(0.1066)	0.035
	Huber_Mcp	0.2128(0.0533)	0.006	0.3506(0.0700)	0.001	0.5857(0.1272)	0.104
	L_2_Lasso	3.9911(2.2393)	0.443	7.5648(3.4787)	1.334	11.7373(5.4510)	2.622
	L_2_Scad	3.9475(2.0684)	0.181	6.5378(3.2249)	0.422	9.3243(4.7560)	0.759
	L_2_Mcp	3.9749(2.1036)	0.164	6.6725(3.3108)	0.333	9.1941(4.5542)	0.270
800	Huber_Lasso	0.2032(0.0488)	0	0.3396(0.0656)	0.003	0.5921(0.1084)	0
	Huber_Scad	0.1888(0.0466)	0	0.2999(0.0542)	0.003	0.4535(0.0724)	0
	Huber_Mcp	0.1879(0.0462)	0	0.3010(0.0539)	0.003	0.4530(0.0718)	0
	L_2_Lasso	3.6416(1.8760)	0.402	6.3944(2.9109)	1.198	9.9321(4.5436)	2.396
	L_2_Scad	3.5713(1.7869)	0.138	5.6202(2.6785)	0.385	7.8234(3.9427)	0.615
	L_2_Mcp	3.6292(1.8216)	0.148	5.8118(2.6989)	0.340	7.7410(3.8194)	0.340
1000	Huber_Lasso	0.1901(0.0435)	0	0.3198(0.0586)	0	0.5212(0.0909)	0
	Huber_Scad	0.1761(0.0424)	0	0.2783(0.0480)	0	0.4156(0.0614)	0
	Huber_Mcp	0.1772(0.0421)	0	0.2787(0.0472)	0	0.4140(0.0599)	0
	L_2_Lasso	3.3961(1.5837)	0.343	6.0246(3.0721)	1.145	9.2563(4.1259)	2.286
	L_2_Scad	3.2695(1.4331)	0.108	5.4273(2.7061)	0.444	7.2424(3.5309)	0.558
	L_2_Mcp	3.3161(1.4947)	0.177	5.5705(2.8340)	0.529	7.2207(3.5177)	0.342
2000	Huber_Lasso	0.1573(0.0408)	0	0.2557(0.0446)	0	0.4033(0.0530)	0
	Huber_Scad	0.1459(0.0332)	0	0.2291(0.0358)	0	0.3267(0.0401)	0
	Huber_Mcp	0.1446(0.0342)	0	0.2291(0.0358)	0	0.3262(0.0410)	0
	L_2_Lasso	2.5780(1.1761)	0.244	4.5743(1.8786)	0.845	7.1840(2.9250)	1.959
	L_2_Scad	2.5260(1.0939)	0.098	4.0856(1.6650)	0.381	5.8326(2.4949)	0.629
	L_2_Mcp	2.5780(1.1070)	0.123	4.2698(1.6642)	0.592	5.9576(2.7359)	0.728
4000	Huber_Lasso	0.1300(0.0272)	0	0.2157(0.0345)	0	0.3321(0.0382)	0
	Huber_Scad	0.1238(0.0266)	0	0.1935(0.0282)	0	0.2715(0.0300)	0
	Huber_Mcp	0.1239(0.0266)	0	0.1933(0.0283)	0	0.2715(0.0301)	0
	L_2_Lasso	1.9665(0.8091)	0.130	3.4642(1.4143)	0.621	5.5383(1.8684)	1.629
	L_2_Scad	1.9958(0.7870)	0.055	3.1574(1.1429)	0.299	4.7194(1.6387)	0.912
	L_2_Mcp	2.0231(0.7878)	0.108	3.3470(1.1299)	0.577	5.0823(1.7436)	1.383

To eliminate any potential influence of different scales among the variables, both the response variable and the covariates are centralized and standardized, ensuring they all have zero means and unit standard deviations. To facilitate comparative analysis, the final trimester of data is designated as the test dataset, comprising a sample size of 73. In contrast, the remaining data are assigned as the training data set. The correlation analysis of the vectorized covariates, as depicted in Fig. 2, demonstrates a pronounced correlation among the matrix covariates. It is worth noting that if the matrix predictor is transformed into a vectorized form, its valuable natural structure information may be lost.

The tuning parameters for all methods were chosen by using the ten-fold cross-validation method. Fig. 3 depicts the estimator under different methods. The figure indicates that PM2.5 is influenced not only by individual indicators but also by their interactions. For example, the estimator $\hat{\Theta}$ derived from the Huber method with SCAD penalty reveals that the variables NO_2 (column 2), $NO_2 \times TEMP$ (column 12), $NO_2 \times pres$ (column 13), $NO_2 \times DEWP$ (column 14) and $NO_2 \times WSPM$ (column 15) are all significant.

Note that when heavy-tailed data are present, one often employs tail index estimation in extreme value theory, such as the widely used Hill estimator in [14] to check whether heavy tails exist. To further motivate robustness against possible heavy tails, we employ the Hill estimator to estimate the tail indexes for residuals of the trace regression models studied in this paper. As demonstrated in Fig. 5, Hill estimates indicate that residuals may have infinite variance. The estimates of the probability density function of the residuals, depicted in Fig. 4, further reveal a slightly right-skewed distribution.Table 5 shows that both the prediction mean square error (PMSE) $= \frac{1}{73}\sum_{i=1}^{73}(Y_i - \hat{Y}_i)^2$

Table 2
Simulation results for Scenario 2.

n	Methods	$d=4$		$d=8$		$d=15$	
		ERR	D_{rk}	ERR	D_{rk}	ERR	D_{rk}
500	Huber_Lasso	0.2328(0.0539)	0.001	0.3837(0.0600)	0	0.8459(0.1296)	0
	Huber_Scad	0.2226(0.0513)	0	0.3756(0.0582)	0	0.5403(0.0852)	0
	Huber_Mcp	0.2227(0.0492)	0	0.3760(0.0589)	0	0.5969(0.1246)	0
	L_2_Lasso	0.7429(0.4135)	0.019	1.1732(0.7733)	0.121	2.0259(1.0146)	0.341
	L_2_Scad	0.7473(0.4137)	0.012	1.1206(0.6599)	0.024	1.7099(0.8036)	0.030
	L_2_Mcp	0.7473(0.4220)	0.013	1.1192(0.6704)	0.016	1.7029(0.7939)	0.017
800	Huber_Lasso	0.1909(0.0396)	0	0.2962(0.0476)	0	0.4946(0.0559)	0
	Huber_Scad	0.1830(0.0374)	0	0.2840(0.0431)	0	0.4738(0.0600)	0
	Huber_Mcp	0.1835(0.0372)	0	0.2848(0.0434)	0	0.4961(0.0594)	0
	L_2_Lasso	0.6555(0.3588)	0.021	0.9750(0.4906)	0.096	1.6265(0.8613)	0.264
	L_2_Scad	0.6586(0.3733)	0.014	0.9555(0.4746)	0.020	1.4812(0.6996)	0.041
	L_2_Mcp	0.6587(0.3733)	0.014	0.9577(0.4767)	0.016	1.4823(0.6918)	0.011
1000	Huber_Lasso	0.1696(0.0440)	0.004	0.2680(0.0445)	0	0.4214(0.0438)	0
	Huber_Scad	0.1622(0.0377)	0.004	0.2547(0.0408)	0	0.4088(0.0470)	0
	Huber_Mcp	0.1622(0.0377)	0.004	0.2550(0.0395)	0	0.4180(0.0454)	0
	L_2_Lasso	0.5896(0.2891)	0.027	0.9191(0.5420)	0.079	1.4242(0.7344)	0.203
	L_2_Scad	0.5934(0.2962)	0.006	0.9050(0.5093)	0.011	1.3477(0.5794)	0.029
	L_2_Mcp	0.5934(0.2962)	0.004	0.9048(0.5116)	0.003	1.3427(0.5870)	0.012
2000	Huber_Lasso	0.1314(0.0318)	0	0.2012(0.0300)	0	0.2969(0.0306)	0
	Huber_Scad	0.1244(0.0266)	0	0.1908(0.0252)	0	0.2818(0.0278)	0
	Huber_Mcp	0.1241(0.0267)	0	0.1908(0.0254)	0	0.2817(0.0286)	0
	L_2_Lasso	0.4914(0.2572)	0.025	0.7233(0.3625)	0.064	1.1175(0.5405)	0.145
	L_2_Scad	0.4955(0.2695)	0.006	0.7195(0.3568)	0.015	1.0731(0.4848)	0.021
	L_2_Mcp	0.4955(0.2695)	0.004	0.7195(0.3583)	0.009	1.0748(0.4844)	0.019
4000	Huber_Lasso	0.1013(0.0225)	0	0.1542(0.0236)	0	0.2257(0.0238)	0
	Huber_Scad	0.0962(0.0198)	0	0.1464(0.0200)	0	0.2086(0.0202)	0
	Huber_Mcp	0.0962(0.0197)	0	0.1464(0.0197)	0	0.2086(0.0207)	0
	L_2_Lasso	0.3869(0.2167)	0.011	0.5896(0.2761)	0.031	0.8787(0.4029)	0.108
	L_2_Scad	0.3938(0.2229)	0.004	0.5890(0.2790)	0.005	0.8592(0.3939)	0.021
	L_2_Mcp	0.3938(0.2229)	0.003	0.5892(0.2783)	0.002	0.8588(0.3941)	0.014

and the prediction mean absolute deviation (PMAD) $= \frac{1}{73}\sum_{i=1}^{73}|Y_i - \hat{Y}_i|$ of the predictions based on the testing data. It is seen in Table 5 that the results of the PMSEs and PMADs show that our proposed estimates perform better than other methods.

5. Concluding remarks

This study investigates the use of adaptive Huber regression with matrix-type covariates. The concave nuclear norm is selected as the penalty function to estimate the true parameter, as it possesses the desirable oracle property for low-rank structures. Moreover, we utilize an extended local adaptive majorize-minimization algorithm developed in [8] to estimate the coefficient matrix. Based on some assumptions, we establish the statistical error rate of Θ according to the theorem. Furthermore, the efficacy of our method is demonstrated through both simulated data and real data analysis, showcasing its potential in practical applications.

The present study opens up several avenues for future research. Firstly, considering more intricate structures in the parameter matrix Θ in the linear trace model under adaptive Huber loss is a promising direction. For instance, exploring row(column) sparsity [31,37], spline structure [11], and row(column) cluster [15] could lead to more sophisticated models. Secondly, generalizing trace models to account for heavy-tailed or asymmetric errors is an attractive extension. This aspect is under investigation and will be discussed in a separate report. Thirdly, another promising direction is examining tensor data, which contains more intricate structural information compared to matrix-valued data, under adaptive Huber loss. The theoretical analysis is particularly challenging and will be left as

Table 3
Simulation results for Scenario 3.

n	Methods	$d=4$		$d=8$		$d=15$	
		ERR	D_{rk}	ERR	D_{rk}	ERR	D_{rk}
500	Huber_Lasso	0.2833(0.0413)	0	0.4019(0.0498)	0	0.8300(0.1113)	0
	Huber_Scad	0.2806(0.0410)	0	0.3931(0.0426)	0	0.5442(0.0492)	0
	Huber_Mcp	0.2801(0.0407)	0	0.3913(0.0424)	0	0.5498(0.0492)	0
	L_2_Lasso	0.6581(0.3509)	0.019	1.0521(0.5835)	0.083	1.7589(0.9887)	0.264
	L_2_Scad	0.6600(0.3691)	0.001	1.0265(0.5300)	0.019	1.5077(0.7535)	0.030
	L_2_Mcp	0.6610(0.3779)	0	1.0239(0.5321)	0.007	1.5080(0.7515)	0.005
800	Huber_Lasso	0.2058(0.0348)	0.001	0.2966(0.0390)	0	0.4698(0.0512)	0
	Huber_Scad	0.2034(0.0324)	0.001	0.2935(0.0321)	0	0.4035(0.0327)	0
	Huber_Mcp	0.2026(0.0326)	0.001	0.2921(0.0323)	0	0.4073(0.0361)	0
	L_2_Lasso	0.5452(0.3117)	0.017	0.8870(0.4429)	0.051	1.3994(0.7623)	0.188
	L_2_Scad	0.5492(0.3162)	0.004	0.8670(0.4340)	0.017	1.2751(0.6328)	0.032
	L_2_Mcp	0.5499(0.3162)	0.004	0.8673(0.4412)	0.008	1.2751(0.6381)	0.011
1000	Huber_Lasso	0.1701(0.0320)	0	0.2501(0.0292)	0	0.3733(0.0334)	0
	Huber_Scad	0.1667(0.0295)	0	0.2496(0.0263)	0	0.3411(0.0273)	0
	Huber_Mcp	0.1668(0.0298)	0	0.2488(0.0256)	0	0.3425(0.0270)	0
	L_2_Lasso	0.4879(0.2504)	0.016	0.8200(0.4080)	0.062	1.2407(0.6738)	0.164
	L_2_Scad	0.4966(0.2562)	0.005	0.8180(0.4119)	0.013	1.1468(0.5559)	0.026
	L_2_Mcp	0.4965(0.2562)	0.005	0.8179(0.4107)	0.008	1.1459(0.5582)	0.013
2000	Huber_Lasso	0.0822(0.0180)	0	0.1373(0.0188)	0	0.2107(0.0210)	0
	Huber_Scad	0.0763(0.0176)	0	0.1307(0.0175)	0	0.1943(0.0177)	0
	Huber_Mcp	0.0766(0.0175)	0	0.1306(0.0176)	0	0.1942(0.0182)	0
	L_2_Lasso	0.4129(0.1993)	0.007	0.6226(0.2649)	0.019	0.9264(0.4082)	0.072
	L_2_Scad	0.4241(0.2085)	0.003	0.6168(0.2674)	0.005	0.8851(0.3864)	0.010
	L_2_Mcp	0.4241(0.2088)	0.003	0.6167(0.2678)	0.004	0.8846(0.3867)	0.005
4000	Huber_Lasso	0.0579(0.0150)	0	0.0859(0.0115)	0	0.1305(0.0141)	0
	Huber_Scad	0.0513(0.0105)	0	0.0803(0.0103)	0	0.1147(0.0110)	0
	Huber_Mcp	0.0513(0.0104)	0	0.0802(0.0103)	0	0.1147(0.0111)	0
	L_2_Lasso	0.3167(0.1551)	0.005	0.5108(0.2277)	0.024	0.7256(0.3158)	0.078
	L_2_Scad	0.3259(0.1576)	0.001	0.5117(0.2302)	0.006	0.7141(0.3175)	0.009
	L_2_Mcp	0.3259(0.1576)	0.001	0.5117(0.2302)	0.003	0.7138(0.3177)	0.006

future work. Finally, further research is needed to assess the effectiveness of the adaptive Huber trace regression model with low-rank regularization in various applications, such as finance, engineering, and biology. While these topics extend beyond the scope of this article, they will be pursued in future studies.

CRediT authorship contribution statement

Xiangyong Tan: Conceptualization, Formal analysis, Methodology, Software, Writing - original draft. **Ling Peng**: Conceptualization, Formal analysis, Methodology, Writing - review & editing. **Heng Lian**: Conceptualization, Writing - review & editing. **Xiaohui Liu**: Conceptualization, Supervision, Writing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Xiangyong Tan's research was supported by the NSF of China (Grant No. 12201260), China Postdoctoral Science Foundation (Grant No. 2022M711425), Jiangxi Provincial NSF (Grant No.

Table 4
Simulation results for Scenario 4.

n	Methods	$d=4$		$d=8$		$d=15$	
		ERR	D_{rk}	ERR	D_{rk}	ERR	D_{rk}
500	Huber_Lasso	0.1665(0.0339)	0	0.2712(0.0437)	0	0.6778(0.1079)	0
	Huber_Scad	0.1624(0.0347)	0	0.2663(0.0418)	0	0.3840(0.0395)	0
	Huber_Mcp	0.1625(0.0347)	0	0.2663(0.0415)	0	0.3859(0.0416)	0
	L_2_Lasso	0.1484(0.0331)	0	0.2463(0.0397)	0	0.4819(0.0628)	0
	L_2_Scad	0.1483(0.0337)	0	0.2360(0.0354)	0	0.3517(0.0360)	0
	L_2_Mcp	0.1486(0.0336)	0	0.2360(0.0353)	0	0.3501(0.0356)	0
800	Huber_Lasso	0.1270(0.0247)	0	0.2020(0.0278)	0	0.3560(0.0411)	0
	Huber_Scad	0.1259(0.0239)	0	0.2012(0.0287)	0	0.2972(0.0285)	0
	Huber_Mcp	0.1259(0.0240)	0	0.2014(0.0285)	0	0.3061(0.0326)	0
	L_2_Lasso	0.1178(0.0245)	0	0.1929(0.0268)	0	0.3117(0.0331)	0
	L_2_Scad	0.1193(0.0236)	0	0.1880(0.0250)	0	0.2689(0.0254)	0
	L_2_Mcp	0.1193(0.0236)	0	0.1880(0.0249)	0	0.2694(0.0253)	0
1000	Huber_Lasso	0.1108(0.0227)	0	0.1801(0.0214)	0	0.2878(0.0294)	0
	Huber_Scad	0.1094(0.0237)	0	0.1775(0.0225)	0	0.2680(0.0263)	0
	Huber_Mcp	0.1095(0.0237)	0	0.1781(0.0227)	0	0.2728(0.0282)	0
	L_2_Lasso	0.1046(0.0220)	0	0.1701(0.0200)	0	0.2662(0.0259)	0
	L_2_Scad	0.1051(0.0226)	0	0.1681(0.0206)	0	0.2417(0.0229)	0
	L_2_Mcp	0.1051(0.0226)	0	0.1681(0.0206)	0	0.2411(0.0227)	0
2000	Huber_Lasso	0.0729(0.0161)	0	0.1180(0.0149)	0	0.1751(0.0175)	0
	Huber_Scad	0.0736(0.0162)	0	0.1173(0.0148)	0	0.1674(0.0167)	0
	Huber_Mcp	0.0736(0.0162)	0	0.1173(0.0147)	0	0.1674(0.0167)	0
	L_2_Lasso	0.0729(0.0161)	0	0.1180(0.0149)	0	0.1751(0.0175)	0
	L_2_Scad	0.0736(0.0162)	0	0.1173(0.0148)	0	0.1674(0.0167)	0
	L_2_Mcp	0.0736(0.0162)	0	0.1173(0.0147)	0	0.1674(0.0167)	0
4000	Huber_Lasso	0.0523(0.0107)	0	0.0829(0.0107)	0	0.1203(0.0121)	0
	Huber_Scad	0.0527(0.0109)	0	0.0826(0.0107)	0	0.1179(0.0114)	0
	Huber_Mcp	0.0527(0.0109)	0	0.0826(0.0107)	0	0.1178(0.0114)	0
	L_2_Lasso	0.0523(0.0107)	0	0.0829(0.0107)	0	0.1203(0.0121	0
	L_2_Scad	0.0527(0.0109)	0	0.0826(0.0107)	0	0.1179(0.0114)	0
	L_2_Mcp	0.0527(0.0109)	0	0.0826(0.0107)	0	0.1178(0.0114)	0

Table 5
Prediction err based on the testing data.

	Huber_Lasso	Huber_Scad	Huber_Mcp	Ls_Lasso	Ls_Scad	Ls_Mcp	XGBoost	RF
PMSE	0.2055	0.1621	0.1444	0.1955	0.1973	0.1850	0.2548	0.2415
PMAD	0.3029	0.2690	0.2553	0.2818	0.2749	0.2695	0.3215	0.3077

20212BAB211010), Science & Technology research project of the Education Department of Jiangxi Province (Grant No. GJJ2200502) and the Jiangxi Province University Humanities and Social Sciences Project (Grant No. JD22031). Ling Peng's research is supported by the NSF of China (Grant No. 12201259), Jiangxi Provincial NSF (Grant No. 20224BAB211008), and the Science & Technology research project of the Education Department of Jiangxi Province (Grant No. GJJ2200537). Heng Lian's research is supported by the High-level Overseas Talent Program (Grant No. 20223BCJ25017). Xiaohui Liu's research is supported by NSF of China (Grant No. 11971208), the National Social Science Foundation of China (Grant No. 21&ZD152), and the Outstanding Youth Fund Project of the Science and Technology Department of Jiangxi Province (No. 20224ACB211003). This work is also supported by Jiangxi Province Key Laboratory of Data Science in Finance and Economics (Grant No. 2024SSY03201).

Appendix A. Proofs of the main results

In this Appendix, we provide a detailed proof of the main results.

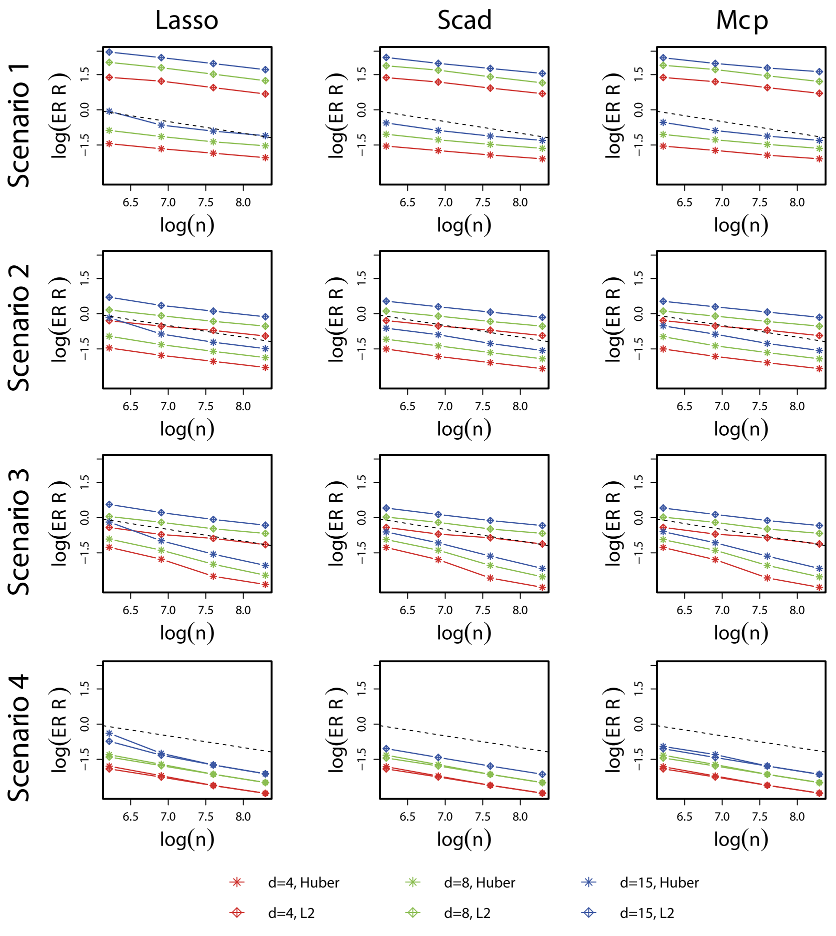

Fig. 1. The averaged statistical error $\|\hat{\Theta}-\Theta^*\|_F$ versus $\log(n)$ for different dimensions d. The slope of the black dashed line is $-1/2$.

Lemma 1. *Suppose Conditions (**C1**)-(**C4**) hold. If $k_l > \eta_-/2$, and $\lambda \geq 2\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}$, we have $\hat{\Delta}_\Theta := \hat{\Theta} - \Theta^* \in \mathcal{C}$.*

Proof. According to the proof of Lemma D2 in [12] and conditioned on $\kappa_l > \eta_-/2$, we have

$$\tilde{L}_{n,\alpha,\lambda}(\Theta) - \tilde{L}_{n,\alpha,\lambda}(\Theta^*) - \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*), \Theta - \Theta^* \rangle \geq (\kappa_l - \frac{\eta_-}{2})\|\Theta - \Theta^*\|_F^2. \tag{16}$$

Thus, by the optimality of (4) and the decomposition in (5), we have

$$\begin{aligned}
0 &\geq \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}) + \lambda\|\hat{\Theta}\|_* - \tilde{L}_{n,\alpha,\lambda}(\Theta^*) - \lambda\|\Theta^*\|_* \\
&\geq -\left|\left\langle \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*), \hat{\Delta}_\Theta \right\rangle\right| + \lambda\left(\|\hat{\Theta}\|_* - \|\Theta^*\|_*\right) \\
&\geq -\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*))\|_{op}\ \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\hat{\Delta}_\Theta)\|_* \\
&\quad - \|\mathcal{P}_{\overline{\mathcal{M}}}(\nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*))\|_{op}\ \|\mathcal{P}_{\overline{\mathcal{M}}}(\hat{\Delta}_\Theta)\|_* + \lambda\left(\|\hat{\Theta}\|_* - \|\Theta^*\|_*\right),
\end{aligned} \tag{17}$$

where the last inequality follows from Hölder's inequality.

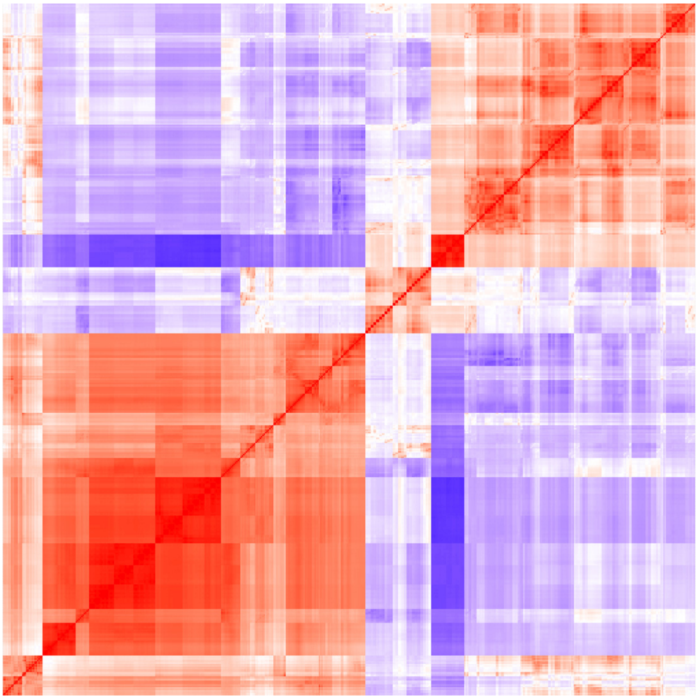

Fig. 2. The correlation Heatmap plot of the predictor matrix, where positive correlation is denoted by red code and negative correlation is indicated by blue code. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

For the third term, it follows from Lemma 3 in [28] that

$$\begin{aligned}\lambda\left(\|\hat{\Theta}\|_* - \|\Theta^*\|_*\right) &\geq \lambda\left(\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\hat{\Delta}_\Theta)\|_* - \|\mathcal{P}_{\overline{\mathcal{M}}}(\hat{\Delta}_\Theta)\|_* - 2\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\Theta^*)\|_*\right)\\ &= \lambda\left(\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\hat{\Delta}_\Theta)\|_* - \|\mathcal{P}_{\overline{\mathcal{M}}}(\hat{\Delta}_\Theta)\|_*\right).\end{aligned} \tag{18}$$

For the first term in the right-hand side of inequality (17), by the choice of $\lambda \geq 2\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}$ and $\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\nabla \mathcal{Q}(\Theta^*))\|_{op} = 0$, it follows that

$$\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*))\|_{op} = \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\nabla L_{n,\alpha}(\Theta^*))\|_{op} \leq \frac{1}{2}\lambda. \tag{19}$$

For the second term in the right-hand side of inequality (17), we have that

$$\|\mathcal{P}_{\overline{\mathcal{M}}}(\nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*))\|_{op} \leq \|\mathcal{P}_{\overline{\mathcal{M}}}(\nabla L_{n,\alpha}(\Theta^*))\|_{op} + \lambda \leq \frac{3}{2}\lambda. \tag{20}$$

Combining (17)-(20), we have

$$\begin{aligned}0 &\geq \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}) + \lambda\|\hat{\Theta}\|_* - \tilde{L}_{n,\alpha,\lambda}(\Theta^*) - \lambda\|\Theta^*\|_*\\ &\geq \frac{1}{2}\lambda\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\hat{\Delta}_\Theta)\|_* - \frac{5}{2}\lambda\|\mathcal{P}_{\overline{\mathcal{M}}}(\hat{\Delta}_\Theta)\|_*,\end{aligned}$$

which implies $\hat{\Delta}_\Theta \in \mathcal{C}$. $\square$

Recall that Θ^* denotes the true underlying value of Θ and

$$\Theta^*_\alpha := \arg\min_{\Theta} \mathbb{E} l_\alpha(Y - \langle X, \Theta\rangle).$$

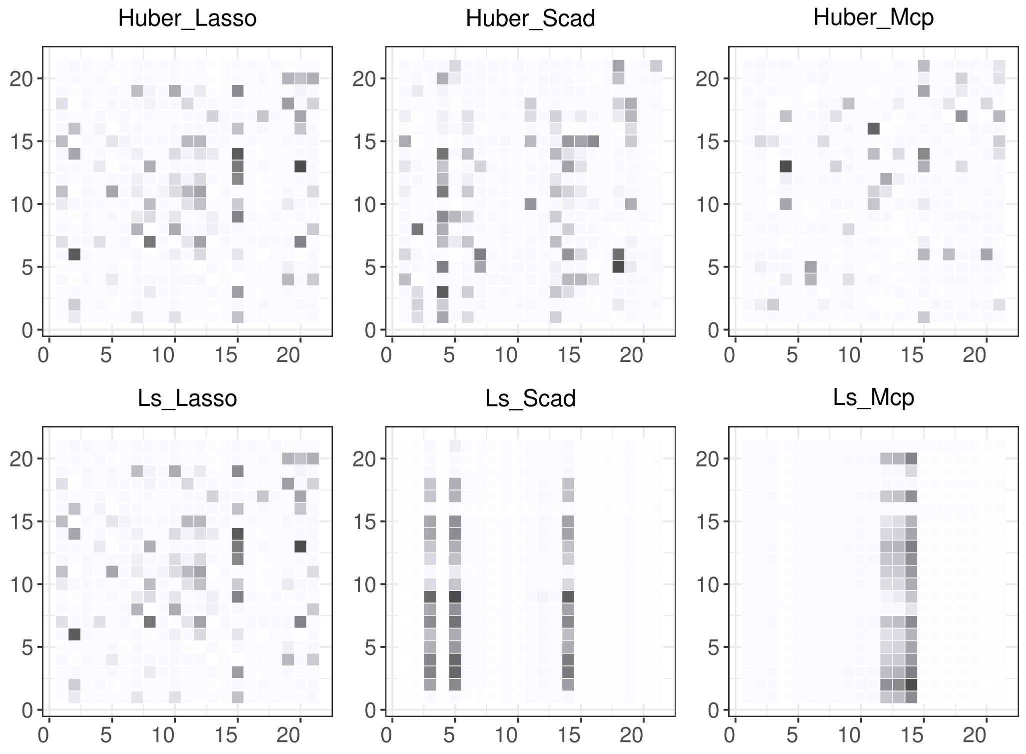

Fig. 3. The estimates of Beijing Air Quality data.

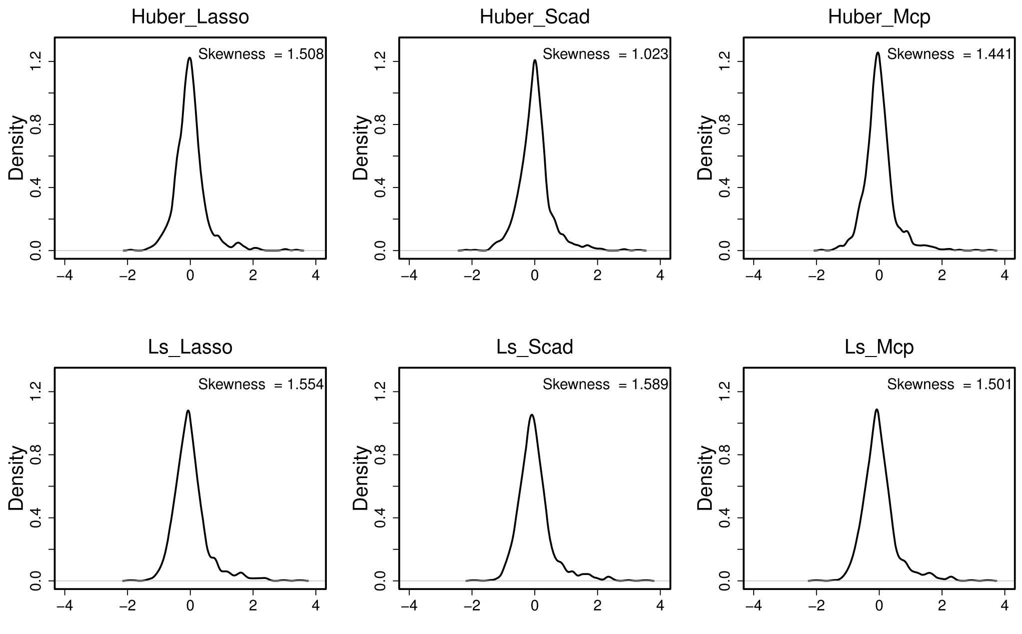

Fig. 4. The density function of the residuals from the six methods.

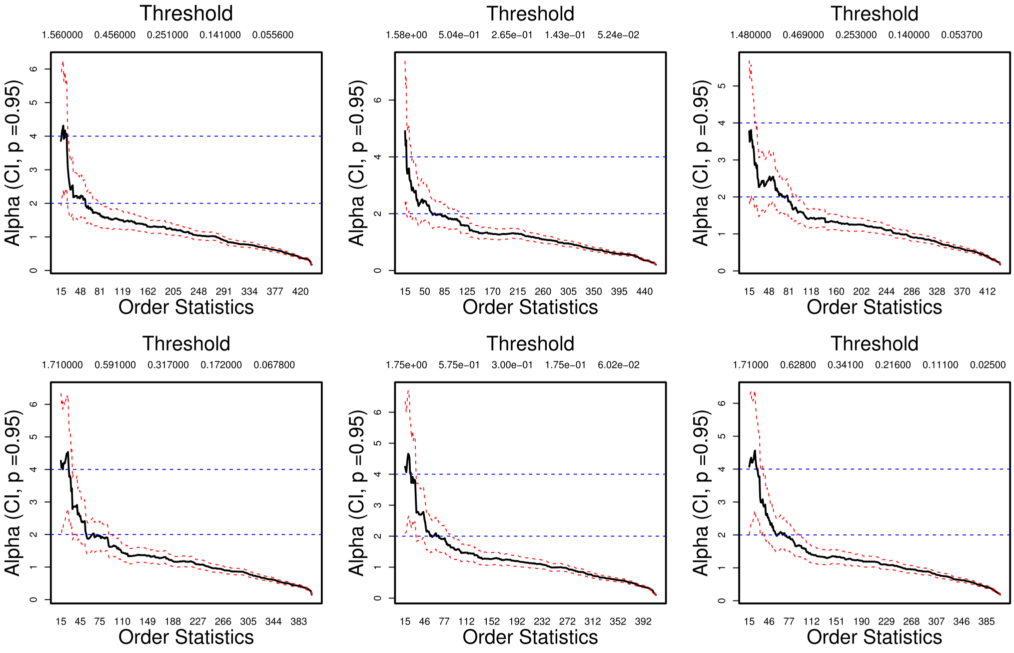

Fig. 5. Hill-plots of the residuals from the six methods. The first row of the figure depicts the methods of *Huber_Lasso*, *Huber_Scad*, and *Huber_Mcp*. The second row illustrates the methods of *Ls_Lasso*, *Ls_Scad*, and *Ls_Mcp*.

Proof of Proposition 1. Let $l(x)=x^2$ be the quadratic loss function. Since Θ_α^* is the minimizer of $\mathbb{E}l_\alpha(Y-\langle X,\Theta\rangle)$, it follows that

$$
\begin{aligned}
&\mathbb{E}\left[l(Y-\langle X,\Theta_\alpha^*\rangle)-l(Y-\langle X,\Theta^*\rangle)\right]\\
=&\mathbb{E}\left[l(Y-\langle X,\Theta_\alpha^*\rangle)-l_\alpha(Y-\langle X,\Theta_\alpha^*\rangle)\right]+\mathbb{E}\left[l_\alpha(Y-\langle X,\Theta_\alpha^*\rangle)-l_\alpha(Y-\langle X,\Theta^*\rangle)\right]\\
&+\mathbb{E}\left[l_\alpha(Y-\langle X,\Theta^*\rangle)-l(Y-\langle X,\Theta^*\rangle)\right]\\
\le&\mathbb{E}\left[(l-l_\alpha)(Y-\langle X,\Theta_\alpha^*\rangle)\right]-\mathbb{E}\left[(l-l_\alpha)(Y-\langle X,\Theta^*\rangle)\right],
\end{aligned}
\tag{21}
$$

where $(l-l_\alpha)(x)=(|x|-\alpha)^2I_{\{|x|>\alpha\}}$. Thus, through Taylor's expansion, we can get that

$$
\begin{aligned}
&\mathbb{E}\left[(l-l_\alpha)(Y-\langle X,\Theta_\alpha^*\rangle)\right]-\mathbb{E}\left[(l-l_\alpha)(Y-\langle X,\Theta^*\rangle)\right]\\
\le&2\mathbb{E}\left[\left(|Y-\langle X,\tilde\Theta\rangle|-\alpha\right)I_{\{|Y-\langle X,\tilde\Theta\rangle|>\alpha\}}\left|\langle X,\Theta_\alpha^*-\Theta^*\rangle\right|\right],
\end{aligned}
$$

where $\tilde\Theta=\kappa\Theta_\alpha^*+(1-\kappa)\Theta^*$ with κ being some constant lying between 0 and 1. Denote the distribution and expectation of ε conditioning on X as $\mathbb{P}_{\varepsilon|X}$ and $\mathbb{E}_{\varepsilon|X}$, respectively, we have

$$
\begin{aligned}
&\mathbb{E}_{\varepsilon|X}\left[\left(|Y-\langle X,\tilde\Theta\rangle|-\alpha\right)I_{\{|Y-\langle X,\tilde\Theta\rangle|>\alpha\}}\right]\\
=&\int_0^\infty\mathbb{P}_{\varepsilon|X}\left[\left(|Y-\langle X,\tilde\Theta\rangle|-\alpha\right)I_{\{|Y-\langle X,\tilde\Theta\rangle|>\alpha\}}>t\right]dt\\
=&\int_\alpha^\infty P_{\varepsilon|X}\left(|Y-\langle X,\tilde\Theta\rangle|>t\right)dt\le\int_\alpha^\infty\frac{\mathbb{E}_{\varepsilon|X}|Y-\langle X,\tilde\Theta\rangle|^{1+\delta}}{t^{1+\delta}}dt
\end{aligned}
$$

$$\leq \frac{1}{\delta}\alpha^{-\delta}\mathbb{E}_{\varepsilon|X}|Y-\langle X,\tilde{\Theta}\rangle|^{1+\delta}.$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[(l-l_\alpha)(Y-\langle X,\Theta_\alpha^*\rangle)\right]-\mathbb{E}\left[(l-l_\alpha)(Y-\langle X,\Theta^*\rangle)\right]\\
\leq& 2\frac{1}{\delta}\alpha^{-\delta}\mathbb{E}\left[|Y-\langle X,\tilde{\Theta}\rangle|^{1+\delta}\left|\langle X,\Theta_\alpha^*-\Theta^*\rangle\right|\right]\\
=& 2\frac{1}{\delta}\alpha^{-\delta}\mathbb{E}\left[|\varepsilon+\langle X,\Theta^*-\tilde{\Theta}\rangle|^{1+\delta}\left|\langle X,\Theta_\alpha^*-\Theta^*\rangle\right|\right]\\
\leq& \frac{1}{\delta}2^{k+1}\alpha^{-\delta}\left(\mathbb{E}\left[|\varepsilon|^{1+\delta}|\langle X,\Theta^*-\tilde{\Theta}\rangle|\right]+\mathbb{E}\left[|\langle X,\Theta^*-\tilde{\Theta}\rangle|^{1+\delta}\left|\langle X,\Theta_\alpha^*-\Theta^*\rangle\right|\right]\right).
\end{aligned} \tag{22}$$

By the Cauchy-Schwarz inequality and conditions (**C1**)-(**C2**), we have

$$\begin{aligned}
\mathbb{E}\left[|\varepsilon|^{1+\delta}|\langle X,\Theta^*-\tilde{\Theta}\rangle|\right]&=\mathbb{E}\left[\mathbb{E}_{\varepsilon|X}\left[|\varepsilon|^{1+\delta}\right]|\langle X,\Theta^*-\tilde{\Theta}\rangle|\right]\\
&\leq\left[\mathbb{E}\left[E_{\varepsilon|X}\left[|\varepsilon|^{1+\delta}\right]\right]^2\right]^{1/2}\left[\mathbb{E}|\langle X,\Theta^*-\tilde{\Theta}\rangle|^2\right]^{1/2}\\
&\leq K_\varepsilon^{1/2}\left[\mathbb{E}|(vec(\Theta_\alpha^*-\tilde{\Theta}))^\top vec(X)(vec(X))^\top vec(\Theta_\alpha^*-\tilde{\Theta})|\right]^{1/2}\\
&\leq (K_\varepsilon\rho_u)^{1/2}\|\Theta_\alpha^*-\tilde{\Theta}\|_F,
\end{aligned} \tag{23}$$

and by the condition (**C3**), we have $\mathbb{E}\left[\exp(t(vec(X)^T)u)\right]\leq\exp\left(ct^2K_X^2\|u\|^2\right)$ for any $u\in\mathbb{R}^{d_1d_2}$, where c is a constant independent of u. Then $(vec(X))^\top vec(\Theta^*-\tilde{\Theta})$ is sub-Gaussian with the $2(1+\delta)$-th moment bounded by $C^2K_X^2$ where C depends only on $1+\delta$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[|\langle X,\Theta^*-\tilde{\Theta}\rangle|^{1+\delta}\left|\langle X,\Theta_\alpha^*-\Theta^*\rangle\right|\right]&\leq\left[\mathbb{E}|\langle X,\Theta^*-\tilde{\Theta}\rangle|^{2(1+\delta)}\right]^{1/2}\left[\mathbb{E}\left|\langle X,\Theta_\alpha^*-\Theta^*\rangle\right|^2\right]^{1/2}\\
&\leq C\rho_u^{1/2}K_X^{1+\delta}\|\Theta_\alpha^*-\Theta^*\|_F,
\end{aligned} \tag{24}$$

where C is some absolute constant. Therefore, the inequalities (21)-(24) combined give the upper bound

$$\mathbb{E}\left[l(Y-\langle X,\Theta_\alpha^*\rangle)-l(Y-\langle X,\Theta^*\rangle)\right]\leq C\rho_u^{1/2}(K_\varepsilon^{1/2}+K_X^{1+\delta})\alpha^{-\delta}\|\Theta_\alpha^*-\Theta^*\|_F.$$

It is worth noting that from condition (**C2**), we have

$$\begin{aligned}
&\mathbb{E}\left[l(Y-\langle X,\Theta_\alpha^*\rangle)-l(Y-\langle X,\Theta^*\rangle)\right]\\
&\quad=\mathbb{E}\left\{\langle X,\Theta_\alpha^*-\Theta^*\rangle^2\right\}\\
&\quad\geq\rho_l\|\Theta_\alpha^*-\Theta^*\|_F^2.
\end{aligned}$$

The above two results complete the proof of Proposition 1. □

Proof of Proposition 2. For any $\Theta\in B^*(R)\cap\mathcal{C}$, note that

$$\begin{aligned}
&\langle\nabla L_{n,\alpha,\lambda}(\Theta)-\nabla L_{n,\alpha,\lambda}(\Theta^*),\Theta-\Theta^*\rangle\\
=&\frac{1}{n}\sum_{i=1}^n\left(l'_\alpha(y_i-\langle X_i,\Theta^*\rangle)-l'_\alpha(y_i-\langle X_i,\Theta\rangle)\right)\langle X_i,\Theta-\Theta^*\rangle\\
=&\frac{1}{n}\sum_{i=1}^n\left(l'_\alpha(\varepsilon_i)-l'_\alpha(y_i-\langle X_i,\Theta\rangle)\right)\langle X_i,\Theta-\Theta^*\rangle.
\end{aligned}$$

Denote $\Delta_\Theta = \Theta - \Theta^*$ and the following event

$$E_i = \{|\varepsilon_i| \le \alpha/2\} \cap \{|\langle X_i, \Delta_\Theta\rangle| \le \alpha \|\Sigma^{1/2} vec(\Delta_\Theta)\|/(2R)\}.$$

For all $\Theta \in \mathcal{B}^*(R) \cap \mathcal{C}$, on E_i, it holds that $|y_i - \langle X_i, \Theta\rangle| \le |\varepsilon_i| + |\langle X_i, \Delta_\Theta\rangle| \le \alpha$. Note that $l''_\alpha(x) = 2$ for all $|x| \le \alpha$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(l'_\alpha(\varepsilon_i) - l'_\alpha(y_i - \langle X_i, \Theta\rangle)\right)\langle X_i, \Delta_\Theta\rangle \\
\ge&\frac{1}{n}\sum_{i=1}^n \left(l'_\alpha(\varepsilon_i)\rangle - l'_\alpha(\varepsilon_i - \langle X_i, \Delta_\Theta\rangle)\right)\langle X_i, \Delta_\Theta\rangle I\{E_i\} \\
\ge&\frac{2}{n}\sum_{i=1}^n \langle X_i, \Delta_\Theta\rangle^2 I\left\{|\langle X_i, \Delta_\Theta\rangle| \le \alpha \|\Sigma^{1/2} vec(\Delta_\Theta)\|/2R\right\} I\left\{|\varepsilon_i| \le \alpha/2\right\} \\
\ge&\frac{2}{n}\sum_{i=1}^n \psi_{\alpha\|\Sigma^{1/2} vec(\Delta_\Theta)\|/(2R)}(\langle X_i, \Delta_\Theta\rangle) I\left\{|\varepsilon_i| \le \alpha/2\right\},
\end{aligned}$$

where for a truncation level $u > 0$, $\psi_u(x)$ is defined as

$$\psi_u(x) = \begin{cases} x^2, & \text{for } |x| \le u/2, \\ (x-u)^2, & \text{for } u/2 < x \le u, \\ (x+u)^2, & \text{for } -u < x \le -u/2, \\ 0, & \text{for } |x| > u. \end{cases}$$

By construction, $\psi_u(x)$ is u-Lipschitz and satisfies

$$x^2 I\{|x| \le u/2\} \le \psi_u(x) \le x^2 I\{|x| \le u\}. \tag{25}$$

Denote

$$G(\Theta) := \frac{1}{n}\sum_{i=1}^n \psi_{\alpha\|\Sigma^{1/2} vec(\Theta-\Theta^*)\|/(2R)}(\langle X_i, \Theta - \Theta^*\rangle) I\left\{|\varepsilon_i| \le \alpha/2\right\}.$$

Therefore, by combining the above inequalities, we have

$$\langle \nabla L_{n,\alpha,\lambda}(\Theta) - \nabla L_{n,\alpha,\lambda}(\Theta^*), \Theta - \Theta^*\rangle \ge G(\Theta).$$

For $R > 0$, define the following empirical process

$$\mathcal{V}(R) := \sup_{\Theta \in \mathcal{B}^*(R) \cap \mathcal{C}} \frac{|G(\Theta) - \mathbb{E}[G(\Theta^*)]|}{\|\Sigma^{1/2} vec(\Theta - \Theta^*)\|^2}.$$

It can be easily seen that

$$\frac{\langle \nabla L_{n,\alpha,\lambda}(\Theta) - \nabla L_{n,\alpha,\lambda}(\Theta^*), \Theta - \Theta^*\rangle}{\|\Sigma^{1/2} vec(\Theta - \Theta^*)\|^2} \ge \frac{\mathbb{E}[G(\Theta^*)]}{\|\Sigma^{1/2} vec(\Theta - \Theta^*)\|^2} - \mathcal{V}(R). \tag{26}$$

To obtain the lower bound of the right hand of the above inequality, we need to establish the upper bound for $\mathcal{V}(R)$ and the lower bound for $\mathbb{E}[G(\Theta^*)]$, respectively.

By the Markov's inequality, we have

$$\mathbb{E}[G(\Theta)] \ge \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \langle X_i, \Delta_\Theta\rangle^2 I\{|\langle X_i, \Delta_\Theta\rangle| \le \alpha \|\Sigma^{1/2} vec(\Delta_\Theta)\|/(4R)\} I\left\{|\varepsilon_i| \le \alpha/2\right\}\right]$$

$$
\begin{aligned}
&\geq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\langle X_i, \Delta_\Theta\rangle^2 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle X_i, \Delta_\Theta\rangle^2 I\{|\langle X_i, \Delta_\Theta\rangle| \geq \alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(4R)\}\right] \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle X_i, \Delta_\Theta\rangle^2 I\left\{|\varepsilon_i| \leq \alpha/2\right\}\right] \\
&\geq \|\Sigma^{1/2}vec(\Delta_\Theta)\|^2 - \sqrt{C}(4R/\alpha)^2\|\Sigma^{1/2}vec(\Delta_\Theta)\|^{-2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\langle X_i, \Delta_\Theta\rangle^4 \\
&\qquad - (2/\alpha)^{1+\delta}\frac{1}{n}\sum_{i=1}^{n}\nu_{i,\delta}\mathbb{E}\langle X_i, \Delta_\Theta\rangle^2 \\
&\geq \|\Sigma^{1/2}vec(\Delta_\Theta)\|^2\left(1 - \sqrt{C}(4R/\alpha)^2 - \nu_\delta(2/\alpha)^{1+\delta}\right).
\end{aligned}
$$

Take $\alpha \geq 2\max\{4C^{1/4}R, 2\nu_\delta^{1/(1+\delta)}\}$, we have

$$
E[G(\Theta)] \geq \frac{1}{2}\|\Sigma^{1/2}vec(\Theta - \Theta^*)\|^2.
$$

For $\mathcal{V}(R)$, note that $0 \leq \psi_u(x) \leq u^4/4$, Denote $G(\Theta) - \mathbb{E}[G(\Theta)] = \frac{1}{n}\sum_{i=1}^n H_i(\Theta)$, we have

$$
0 \leq \frac{H_i(\Theta)}{\|\Sigma^{1/2}vec(\Theta - \Theta^*)\|^2} \leq \frac{\alpha^2}{16R^2}.
$$

Besides, according to (25), we have

$$
\begin{aligned}
\sigma_n^2 &:= \frac{1}{n}\sum_{i=1}^{n}\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}}\mathbb{E}\left[\frac{H_i^2(\Theta)}{\|\Sigma^{1/2}vec(\Theta - \Theta^*)\|^4}\right] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}}\frac{E\left[\psi^2_{\alpha\|\Sigma^{1/2}vec(\Theta-\Theta^*)\|/(2R)}(\langle X_i, \Theta - \Theta^*\rangle)\right]}{\|\Sigma^{1/2}vec(\Theta - \Theta^*)\|^4} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}}\frac{E\langle X_i, \Theta - \Theta^*\rangle^4}{\|\Sigma^{1/2}vec(\Theta - \Theta^*)\|^4} \\
&\leq C.
\end{aligned}
$$

Therefore, by the functional version of Bennett's inequality (Theorem 7.3 in [1]), for any $t > 0$, we have

$$
\mathcal{V}(R) \leq \mathbb{E}\mathcal{V}(R) + \left(\sqrt{2}C + \frac{\alpha}{2R}\sqrt{\mathbb{E}\mathcal{V}(R)}\right)\sqrt{\frac{t}{n}} + \frac{\alpha^2}{16R^2}\left(\frac{t}{3n}\right), \tag{27}
$$

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

In the following, we need to bound the expected value $\mathbb{E}\mathcal{V}(R)$. By the standard symmetrization argument (Lemma 2.3.6 in [32]) and the connection between the Rademacher complexity $\mathcal{R}_n(\Theta)$ and the Gaussian complexity $\mathcal{G}_n(\Theta)$ (Lemma 4.5 in [20]), we have

$$
\mathbb{E}\mathcal{V}(R) \leq 2\mathcal{R}_n(\Theta) \leq 2\sqrt{\frac{\pi}{2}}\mathcal{G}_n(\Theta). \tag{28}
$$

Here, the Gaussian complexity is denoted as

$$
\mathcal{G}_n(\Theta) = \mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n}z_i\frac{\psi_{\alpha\|\Sigma^{1/2}vec(\Theta-\Theta^*)\|/(2R)}(\langle X_i, \Theta - \Theta^*\rangle)I\left\{|\varepsilon_i| \leq \alpha/2\right\}}{\|\Sigma^{1/2}vec(\Theta - \Theta^*)\|^2}\right|\right],
$$

where $z_i \overset{i.i.d.}{\sim} N(0,1)$ and independent of y_i and X_i. Conditioned on $\{X_i\}_{i=1}^n$, define the following Gaussian process

$$Z_\Theta := \frac{1}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|^2}\cdot\frac{1}{n}\sum_{i=1}^n z_i\left(\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}(\langle X_i,\Delta_\Theta\rangle)I\{|\varepsilon_i|\le\alpha/2\}\right).$$

Then, the Gaussian complexity can be rewritten as

$$\mathcal{G}_n(\Theta)=\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}}|Z_\Theta|\right]. \tag{29}$$

Note that for $\Theta,\tilde{\Theta}\in\mathcal{B}^*(R)\cap\mathcal{C}$, we have

$$\begin{aligned}
&\mathrm{Var}\left(Z_\Theta-Z_{\tilde{\Theta}}\right)\\
&\le\frac{1}{n^2}\sum_{i=1}^n(I\{|\varepsilon_i|\le\alpha/2\})^2\\
&\left(\frac{\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}(\langle X_i,\Delta_\Theta\rangle)}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|^2}-\frac{\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|/(2R)}(\langle X_i,\Delta_{\tilde{\Theta}}\rangle)}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|^2}\right)^2\\
&\le\frac{1}{n^2}\sum_{i=1}^n\left(\frac{\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}(\langle X_i,\Delta_\Theta\rangle)}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|^2}-\frac{\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|/(2R)}(\langle X_i,\Delta_{\tilde{\Theta}}\rangle)}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|^2}\right)^2
\end{aligned}$$

Using the homogeneity property $\frac{1}{c^2}\psi_{cu}(cx)=\psi_u(x),\ \forall c>0$, and the fact that ψ_u is u-Lipschitz, we have

$$\begin{aligned}
&\mathrm{Var}\left(Z_\Theta-Z_{\tilde{\Theta}}\right)\\
\le&\frac{1}{n^2}\sum_{i=1}^n\left(\frac{\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}(\langle X_i,\Delta_\Theta\rangle)}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|^2}\right.\\
&\left.-\frac{\frac{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|^2}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|^2}\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}\left(\langle X_i,\Delta_{\tilde{\Theta}}\rangle\cdot\frac{\|\Sigma^{1/2}vec(\Delta_\Theta)\|}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|}\right)}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|^2}\right)^2\\
=&\frac{1}{n^2}\sum_{i=1}^n\frac{1}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|^4}\left(\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}\left(\langle X_i,\Delta_\Theta\rangle\right)\right.\\
&\left.-\psi_{\alpha\|\Sigma^{1/2}vec(\Delta_\Theta)\|/(2R)}\left(\langle X_i,\Delta_{\tilde{\Theta}}\rangle\cdot\frac{\|\Sigma^{1/2}vec(\Delta_\Theta)\|}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|}\right)\right)^2\\
\le&\frac{1}{n^2}\sum_{i=1}^n\frac{\alpha^2}{4R^2\|\Sigma^{1/2}vec(\Delta_\Theta)\|^2}\left(\langle X_i,\Delta_\Theta\rangle-\langle X_i,\Delta_{\tilde{\Theta}}\rangle\cdot\frac{\|\Sigma^{1/2}vec(\Delta_\Theta)\|}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|}\right)^2\\
=&\frac{\alpha^2}{4R^2n^2}\sum_{i=1}^n\left(\frac{\langle X_i,\Delta_\Theta\rangle}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|}-\frac{\langle X_i,\Delta_{\tilde{\Theta}}\rangle}{\|\Sigma^{1/2}vec(\Delta_{\tilde{\Theta}})\|}\right)^2.
\end{aligned}$$

Defining the centered Gaussian process

$$Y_\Theta=\frac{\alpha}{2R\|\Sigma^{1/2}vec(\Delta_\Theta)\|}\cdot\frac{1}{n}\sum_{i=1}^n z_i'\langle X_i,\Delta_\Theta\rangle,$$

where z_i's are independent standard Gaussians and independent of all the previous variables, it follows that

$$\mathrm{Var}\left(Z_\Theta - Z_{\tilde{\Theta}}\right) \le \mathrm{Var}\left(Y_\Theta - Y_{\tilde{\Theta}}\right)$$

Applying the Gaussian comparison inequality (Corollary 3.14 in [20]), we have

$$\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}} Z_\Theta\right] \le 2\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}} Y_\Theta\right]. \tag{30}$$

Note that for any $\Theta_0 \in \mathcal{B}^*(R)\cap\mathcal{C}$, we have

$$\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}} |Z_\Theta|\right] \le \mathbb{E}\left[|Z_{\Theta_0}|\right] + 2\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}} Z_\Theta\right]. \tag{31}$$

Furthermore, note that $0 \le \psi_u(x) \le u^4/4$, we have

$$E\left[|Z_{\Theta_0}|\right] \le \sqrt{\frac{2}{\pi}}\cdot\sqrt{\mathrm{Var}(Z_{\Theta_0})} \le \sqrt{\frac{2}{\pi}}\cdot\sqrt{\mathbb{E}(Z^2_{\Theta_0})} \le \frac{\alpha}{2\sqrt{2\pi}R}\sqrt{\frac{1}{n}}. \tag{32}$$

Finally, for every $\Theta \in \mathcal{B}^*(R)\cap\mathcal{C}$, according to condition ((C2)), we have

$$\|\Sigma^{1/2}vec(\Delta_\Theta)\| = \sqrt{vec^T(\Delta_\Theta)\Sigma vec(\Delta_\Theta)} \ge \rho_l^{1/2}\|\Delta_\Theta\|_F \ge \rho_l^{1/2}r^{-1/2}\|\Delta_\Theta\|_*,$$

which implies

$$\begin{aligned}
\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}} Y_\Theta\right] &= \frac{\alpha}{2R}\mathbb{E}\left[\sup_{\Theta\in\mathcal{B}^*(R)\cap\mathcal{C}} \frac{1}{n}\sum_{i=1}^n z_i' \frac{\langle X_i, \Delta_\Theta\rangle}{\|\Sigma^{1/2}vec(\Delta_\Theta)\|}\right] \\
&\le \frac{\alpha\sqrt{r}}{2R\sqrt{\rho_l}}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n z_i' X_i\right\|_{op}\right] \\
&\le C\frac{\alpha\sqrt{r}}{2R\sqrt{\rho_l}}\sqrt{\frac{d_1+d_2}{n}}.
\end{aligned} \tag{33}$$

Therefore, (28)-(33) combined gives

$$\mathbb{E}\mathcal{V}(R) \le C\sqrt{2\pi}\left(\frac{\alpha}{2\sqrt{2\pi}R}\sqrt{\frac{1}{n}} + \frac{\alpha\sqrt{r}}{2R\sqrt{\rho_l}}\sqrt{\frac{d_1+d_2}{n}}\right)$$

Take $t = d_1 + d_2$ in (27). It can be obtained that with probability at least $1 - e^{-(d_1+d_2)}$, $\mathcal{V}(R) \le \frac{1}{4}$ for sufficiently large n that scales as $\rho_l^{-1}r(\alpha/R)^2(d_1+d_2)$ up to some absolute constant. Therefore, combined with (26) we have

$$\begin{aligned}
\langle \nabla L_{n,\alpha,\lambda}(\Theta) - \nabla L_{n,\alpha,\lambda}(\Theta^*), \Theta - \Theta^*\rangle &\ge \frac{1}{4}\|\Sigma^{1/2}vec(\Theta-\Theta^*)\|^2 \\
&\ge \frac{\rho_l}{4}\|\Theta-\Theta^*\|_F^2,
\end{aligned} \tag{34}$$

uniformly over $\Theta \in \mathcal{B}^*(R)\cap\mathcal{C}$. □

Proof of Theorem 1. Following the proof scheme of Theorem 1 in [9], we first construct a middle point

$$\hat{\Theta}_{t^*} = \Theta^* + t^*(\hat{\Theta} - \Theta^*).$$

We choose $t^* = 1$ for $\|\hat{\Theta} - \Theta^*\|_F \le R$ and $t^* = R/\|\hat{\Theta} - \Theta^*\|_F$ for $\|\hat{\Theta} - \Theta^*\|_F > R$. Therefore $\|\hat{\Theta}_{t^*} - \Theta^*\|_F \le R$. Denote $\hat{\Delta}_{\Theta,t^*} = \hat{\Theta}_{t^*} - \Theta^*$. According to Lemma 1, for the choice of $\lambda \ge 2\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}$, we have $\hat{\Delta}_\Theta \in \mathcal{C}$. Since $\hat{\Delta}_{\Theta,t^*}$ is parallel to $\hat{\Delta}_\Theta$, $\hat{\Delta}_{\Theta,t^*}$ also falls in this cone.

According to the proof of Lemma D1 in [12] and RSC condition, we have

$$\tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}_{t^*}) - \tilde{L}_{n,\alpha,\lambda}(\Theta^*) - \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*), \hat{\Theta}_{t^*} - \Theta^* \rangle \ge (\kappa_l - \frac{\eta_-}{2})\|\hat{\Theta}_{t^*} - \Theta^*\|_F^2,$$
$$\tilde{L}_{n,\alpha,\lambda}(\Theta^*) - \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}_{t^*}) - \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}_{t^*}), \Theta^* - \hat{\Theta}_{t^*} \rangle \ge (\kappa_l - \frac{\eta_-}{2})\|\Theta^* - \hat{\Theta}_{t^*}\|_F^2.$$

Adding the above two inequalities implies

$$(2\kappa_l - \eta_-)\|\hat{\Delta}_{\Theta,t^*}\|_F^2 \le \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}_{t^*}) - \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*), \hat{\Delta}_{\Theta,t^*} \rangle =: D_L^s(\hat{\Theta}_{t^*}, \Theta^*), \tag{35}$$

where $D_L^s(\cdot)$ is the symmetric Bregman divergence. By Lemma C.1. of [30], $D_L^s(\hat{\Theta}_{t^*}, \Theta^*) \le t^* D_L^s(\hat{\Theta}, \Theta^*)$. It follows that

$$(2\kappa_l - \eta_-)\|\hat{\Delta}_{\Theta,t^*}\|_F^2 \le t^* D_L^s(\hat{\Theta}, \Theta^*) = \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}) - \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*), \hat{\Delta}_{\Theta,t^*} \rangle.$$

Since $\hat{\Theta}$ is the minimizer of the optimization problem (6), we shall have the optimality condition $\nabla \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}) + \lambda \hat{G} = 0$ for some subgradient $\hat{G} \in \partial\|\hat{\Theta}\|_*$. By the monotonicity of the subgradient, we have $\langle \hat{G} - G^*, \hat{\Theta} - \Theta^* \rangle \ge 0$, where $G^* \in \partial\|\Theta^*\|_*$. It follows that

$$\begin{aligned}
(2\kappa_l - \eta_-)\|\hat{\Delta}_{\Theta,t^*}\|_F^2 &\le \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\hat{\Theta}) - \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*), \hat{\Delta}_{\Theta,t^*} \rangle \\
&= -\langle \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*) + \lambda \hat{G}, \hat{\Delta}_{\Theta,t^*} \rangle \\
&\le \langle \nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*) + \lambda G^*, \Theta^* - \hat{\Theta}_{t^*} \rangle \\
&\le \left\langle \mathcal{P}_{\overline{\mathcal{M}}^\perp}\left(\nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*) + \lambda G^*\right), \Theta^* - \hat{\Theta}_{t^*} \right\rangle \\
&\quad + \left\langle \mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla \tilde{L}_{n,\alpha,\lambda}(\Theta^*) + \lambda G^*\right), \Theta^* - \hat{\Theta}_{t^*} \right\rangle \\
&:= A_1 + A_2.
\end{aligned} \tag{36}$$

Recall that we have the SVD of $\Theta^* = U\Gamma^* V^\top$, where $\Gamma^* \in \mathbb{R}^{d\times d}$ is the diagonal matrix that contains the nonzero singular values of Θ^* in decreasing order $\sigma_1(\Theta^*) \ge \cdots \ge \sigma_d(\Theta^*) \ge 0$. Define the set $S := \{j \in \{1, \cdots, d\} | \sigma_j(\Theta^*) > 0\}$ and the corresponding complement $S^c := \{j \in \{1, \cdots, d\} | \sigma_j(\Theta^*) = 0\}$.

Step 1. Next, we derive the upper bound of A_1 in (36). Note that the projection $\mathcal{P}_{\overline{\mathcal{M}}^\perp}$ is related to the index set S^c. Denote $\sigma(\Theta^*)$ be the vector of (ordered) singular values of Θ^*, it follows that $\sigma\left(\mathcal{P}_{\overline{\mathcal{M}}^\perp}(\Theta^*)\right) = (\sigma(\Theta^*))_{S^c} = \mathbf{0}$. According to (iii) in assumption (**C4**) that $q_\lambda'(0) = 0$ and the definition of $\mathcal{Q}_\lambda$, we have

$$\mathcal{P}_{\overline{\mathcal{M}}^\perp}\left(\nabla \mathcal{Q}_\lambda(\Theta^*)\right) = \mathcal{P}_{\overline{\mathcal{M}}^\perp}\left(U q_\lambda'(\Gamma^*) V^\top\right) = \mathbf{0}_{d_1\times d_2}.$$

Note that the subgradient of $\|\Theta^*\|_*$ is

$$\partial\|\Theta^*\|_* = \left\{U_r V_r^\top + W^* : \|W^*\|_{op} \le 1, W^* \in \overline{\mathcal{M}}^\perp\right\}. \tag{37}$$

Meanwhile, we have

$$\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\left(\nabla L_{n,\alpha}(\Theta^*)\right)\|_{op} \le \|\nabla L_{n,\alpha}(\Theta^*)\|_{op} \le \lambda.$$

Thus, with the particular choice of $W^* = -\lambda^{-1}\mathcal{P}_{\overline{\mathcal{M}}^\perp}\left(\nabla L_{n,\alpha}(\Theta^*)\right)$ in (37), we have $G^* = U_r V_r^\top + W^* \in \partial\|\Theta^*\|_*$. It follows that

$$\begin{aligned}\mathcal{P}_{\overline{\mathcal{M}}^{\perp}}\left(\nabla\tilde{L}_{n,\alpha,\lambda}(\Theta^*)+\lambda G^*\right)&=\mathcal{P}_{\overline{\mathcal{M}}^{\perp}}\left(\nabla L_{n,\alpha}(\Theta^*)+\nabla\mathcal{Q}_\lambda(\Theta^*)+\lambda G^*\right)\\&=\mathcal{P}_{\overline{\mathcal{M}}^{\perp}}\left(\nabla L_{n,\alpha}(\Theta^*)\right)+\lambda W^*\\&=\mathbf{0}_{d_1\times d_2}.\end{aligned}$$

Therefore, we have

$$A_1=\left\langle\mathcal{P}_{\overline{\mathcal{M}}^{\perp}}\left(\nabla\tilde{L}_{n,\alpha,\lambda}(\Theta^*)+\lambda G^*\right),\Theta^*-\hat{\Theta}_{t^*}\right\rangle=0. \tag{38}$$

Step 2. Then, we consider the upper bound for A_2. Note that

$$\begin{aligned}A_2&=\left\langle\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\tilde{L}_{n,\alpha,\lambda}(\Theta^*)+\lambda G^*\right),\Theta^*-\hat{\Theta}_{t^*}\right\rangle\\&=\left\langle\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla L_{n,\alpha}(\Theta^*)+\nabla\mathcal{R}_\lambda(\Theta^*)\right),\Theta^*-\hat{\Theta}_{t^*}\right\rangle.\end{aligned} \tag{39}$$

Recall that $\mathcal{R}_\lambda(\Theta^*)=\mathcal{Q}_\lambda(\Theta^*)+\lambda\|\Theta^*\|_*$. Projecting $\nabla\mathcal{R}_\lambda(\Theta^*)$ into the subspace $\overline{\mathcal{M}}$ leads to

$$\begin{aligned}\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\mathcal{R}_\lambda(\Theta^*)\right)&=\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\mathcal{Q}_\lambda(\Theta^*)+\lambda G^*\right)\\&=\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\mathcal{Q}_\lambda(\Theta^*)+\lambda(U_rV_r^\top+W^*)\right)\\&=U_r(q_\lambda'(\Gamma_S^*)+\lambda I_{r\times r})V_r^\top,\end{aligned}$$

where $\Gamma_S^*\in\mathbb{R}^{r\times r}$ is the diagonal matrix of singular values corresponding to $j\in S$ and 0 for $j\notin S$. We decompose the index set S of the ordered singular values into two parts:

(i). For $j\in S_1:=\{j\in\{1,\cdots,r\}|\sigma_j(\Theta^*)\geq\gamma\lambda\}$.
(ii). For $j\in S_2:=\{j\in\{1,\cdots,r\}|0<\sigma_j(\Theta^*)<\gamma\lambda\}$.

Note that $S_1\cup S_2=S$. According to (ii) in condition (**C4**) that $q_\lambda'(t)+\lambda=p_\lambda'(t)=0$ for $t\geq\gamma\lambda$, and note that for $\sigma_j(\Theta^*)<\gamma\lambda$, q_λ satisfies the regularity condition (iv) in (**C4**) that $|q_\lambda'(\sigma_j(\Theta^*))|\leq\lambda$. Therefore, $q_\lambda'(\Gamma_S^*)+\lambda I_S\in\mathbb{R}^{r\times r}$ is a diagonal matrix with

$$\left(q_\lambda'(\Gamma_S^*)+\lambda I_S\right)_{jj}=\begin{cases}p_\lambda'(\sigma_j(\Theta^*))=0, & j\in S_1,\\ q_\lambda'(\sigma_j(\Theta^*))+\lambda, & j\in S_2.\end{cases}$$

Due to (iii) and (iv) in condition (**C4**), we have

$$\begin{aligned}\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\mathcal{R}_\lambda(\Theta^*)\right)\right\|_{op}&\leq\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(U_rq_\lambda'(\Gamma_S^*)V_r^\top\right)\right\|_{op}+\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(\lambda U_rV_r^\top\right)\right\|_{op}\\&\leq\max_{j\in S}(q_\lambda'(\Gamma_S^*))_{jj}+\lambda\\&\leq2\lambda.\end{aligned} \tag{40}$$

Moreover, by the choice of λ such that $\lambda\geq2\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}$, (39) combined with (40) lead to

$$\begin{aligned}A_2&=\left\langle\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla L_{n,\alpha}(\Theta^*)\right),\Theta^*-\hat{\Theta}_{t^*}\right\rangle+\left\langle\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\mathcal{R}_\lambda(\Theta^*)\right),\Theta^*-\hat{\Theta}_{t^*}\right\rangle\\&\leq\left(\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla L_{n,\alpha}(\Theta^*)\right)\right\|_{op}+\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(\nabla\mathcal{R}_\lambda(\Theta^*)\right)\right\|_{op}\right)\cdot\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(\Theta^*-\hat{\Theta}_{t^*}\right)\right\|_*\\&\leq\left(\|\nabla L_{n,\alpha}(\Theta^*)\|_{op}+2\lambda\right)\left\|\mathcal{P}_{\overline{\mathcal{M}}}\left(\Theta^*-\hat{\Theta}_{t^*}\right)\right\|_*\\&\leq\frac{5}{2}\lambda\sqrt{2r}\left\|\Theta^*-\hat{\Theta}_{t^*}\right\|_F.\end{aligned} \tag{41}$$

Finally, (36), (38) and (41) combined gives

$$(2\kappa_l - \eta_-)\|\hat{\Delta}_{\Theta,t^*}\|_F^2 \le \frac{5}{2}\lambda\sqrt{2r}\left\|\hat{\Delta}_{\Theta,t^*}\right\|_F,$$

which indicate

$$\|\hat{\Delta}_{\Theta,t^*}\|_F \le \frac{5/2\lambda\sqrt{2r}}{2\kappa_l - \eta_-}.$$

If we choose $R > \frac{5/2\lambda\sqrt{2r}}{2\kappa_l-\eta_-}$ in advance, we have $\hat{\Delta}_{\Theta,t^*} = \hat{\Delta}_{\Theta}$. It follows that

$$\|\hat{\Theta} - \Theta^*\|_F \le \frac{5/2\lambda\sqrt{2r}}{2\kappa_l - \eta_-}.$$

We complete the proof of this theorem. □

Proof of Proposition 3. Note that $\nabla L_{n,\alpha}(\Theta^*) = -\frac{1}{n}\sum_{i=1}^n l'_\alpha(\varepsilon_i)X_i$, where $l'_\alpha(x) = 2sign(x)\min\{|x|,\alpha\}$ for all $x \in \mathbb{R}$. Thus, we have

$$\|\nabla L_{n,\alpha}(\Theta^*)\|_{op} = \max_{u\in\mathcal{S}^{d_1-1}, v\in\mathcal{S}^{d_2-1}} \frac{1}{n}\sum_{i=1}^n l'_\alpha(\varepsilon_i)u^\top X_i v,$$

where $\mathcal{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ be the unit ball in $\mathbb{R}^d$. Now, let $\mathcal{N}^{d_1}$, $\mathcal{N}^{d_1}$ be the 1/3-covering of $\mathcal{S}^{d_1-1}$ and $\mathcal{S}^{d_2-1}$, respectively. For any matrix $B \in \mathbb{R}^{d_1\times d_2}$, define $\omega(B) = \max_{u\in\mathcal{N}^{d_1-1}, v\in\mathcal{N}^{d_2-1}} u^\top B v$. For any given $u \in \mathcal{S}^{d_1-1}$, $v \in \mathcal{S}^{d_2-1}$, there exists $\tilde{u} \in \mathcal{N}^{d_1-1}$, $\tilde{v} \in \mathcal{N}^{d_2-1}$ such that

$$\begin{aligned} u^\top B v &= \tilde{u}^\top B\tilde{v} + \tilde{u}^\top B(v-\tilde{v}) + (u-\tilde{u})^\top B\tilde{v} + (u-\tilde{u})^\top B(v-\tilde{v}) \\ &\le \omega(B) + \frac{7}{9}\|B\|_{op}. \end{aligned}$$

Taking the maximum over all possible u and v, we have

$$\|B\|_{op} = \max_{u\in\mathcal{S}^{d_1-1}, v\in\mathcal{S}^{d_2-1}} u^\top B v \le \omega(B) + \frac{9}{16}\|B\|_{op},$$

which implies $\|B\|_{op} \le \frac{9}{2}\omega(B)$ for any matrix $B \in \mathbb{R}^{d_1\times d_2}$.

For fixed $u \in \mathcal{N}^{d_1-1}$ and $v \in \mathcal{N}^{d_2-1}$, denote $Z_i = u^\top X_i v$, then we have

$$\|\nabla L_{n,\alpha}(\Theta^*)\|_{op} \le \frac{9}{2}\max_{u\in\mathcal{N}^{d_1-1}, v\in\mathcal{N}^{d_2-1}} \frac{1}{n}\sum_{i=1}^n l'_\alpha(\varepsilon_i)Z_i. \tag{42}$$

Recall that under the condition (**C3**), $vec(X_i)$ is sub-Gaussian and $\|Z_i\|_{\psi_2} = \|vec(X_i)(u\otimes v)\|_{\psi_2} = K_X$, where K_X is a constant bounded from zero and infinity.

We first give a bound on $\max_{u\in\mathcal{N}^{d_1-1}, v\in\mathcal{N}^{d_2-1}} P\left(\frac{1}{n}\sum_{i=1}^n l'_\alpha(\varepsilon_i)Z_i > t\right)$. Let us denote $\xi_i = l'_\alpha(\varepsilon_i) = 2sign(\varepsilon_i)\min\{|\varepsilon_i|,\alpha\} \le 2\alpha$, which implies that ξ_i is sub-Gaussian. Let's consider the following centered random vector

$$\frac{1}{n}\sum_{i=1}^n (\xi_i Z_i - \mathbb{E}[\xi_i Z_i]).$$

Under the condition (**C3**), since $vec(X_i)$ is sub-Gaussian, we have Z_i being sub-Gaussian. Then we have $\mathbb{E}|Z_i|^k \le kC^k\Gamma(k/2)$ for all $k \ge 1$ and some positive constant C. Moreover, for the case of $0 < \delta < 1$, note that

$$
\begin{aligned}
E[\xi_i Z_i]^2 \leq 2C^2 E\xi_i^2 &= 8C^2 \mathbb{E}[\min\{|\varepsilon_i|^2, \alpha^2\}] \\
&= 8C^2 \mathbb{E}\left[|\varepsilon_i|^2 I\{|\varepsilon_i| \leq \alpha\} + \alpha^2 I\{|\varepsilon_i| > \alpha\}\right] \\
&\leq 8C^2 \mathbb{E}\left[\alpha^{1-\delta}|\varepsilon_i|^{1+\delta} I\{|\varepsilon_i| \leq \alpha\} + \alpha^{1-\delta}|\varepsilon_i|^{1+\delta} I\{|\varepsilon_i| > \alpha\}\right] \\
&\leq 8C^2 \alpha^{1-\delta} \nu_{i,\delta},
\end{aligned}
$$

hence, we have

$$
\sum_{i=1}^n E[\xi_i Z_i]^2 \leq 8C^2 \alpha^{1-\delta} n \nu_\delta.
$$

It can be easily checked that the Bernstein condition holds, that is,

$$
\begin{aligned}
\sum_{i=1}^n E[\xi_i Z_i]^k \leq \sum_{i=1}^n & E[|\xi_i Z_i|^{k-2} |\xi_i Z_i|^2] \\
&\leq (2\alpha)^{k-2}(k-2)C^{k-2}\Gamma((k-2)/2) \sum_{i=1}^n E[|\xi_i Z_i|^2] \\
&\leq (2\alpha)^{k-2}(k-2)C^{k-2}\Gamma((k-2)/2) \sum_{i=1}^n E[|\xi_i Z_i|^2] \\
&\leq \frac{1}{2} k! (C\alpha)^{k-2} 8C^2 \alpha^{1-\delta} n \nu_\delta,
\end{aligned}
$$

for all $k \geq 3$. Based on Bernstein's inequality, for any $t > 0$, we have

$$
|\frac{1}{n}\sum_{i=1}^n (\xi_i Z_i - \mathbb{E}[\xi_i Z_i])| \geq C\left(\sqrt{\frac{\nu_\delta \alpha^{1-\delta} t}{n}} + \frac{\alpha t}{n}\right),
$$

with probability at most $2e^{-t}$. Recall that $Z_i = u^\top X_i v$ for fixed $u \in \mathcal{N}^{d_1-1}$ and $v \in \mathcal{N}^{d_2-1}$, taking the union bound over $N^{d_1-1} \times \mathcal{N}^{d_2-1}$, combining (42), we have

$$
\|\nabla L_{n,\alpha}(\Theta^*) - \mathbb{E}[\nabla L_{n,\alpha}(\Theta^*)]\|_{op} \leq \frac{9}{2} C\left(\sqrt{\frac{\nu_\delta \alpha^{1-\delta} t}{n}} + \frac{\alpha t}{n}\right) \tag{43}
$$

with probability at least $1 - 2 \cdot 7^{d_1+d_2} \cdot e^{-t}$. It is easy to see that

$$
\|\mathbb{E}[\nabla L_{n,\alpha}(\Theta^*)]\|_{op} \leq \frac{9}{2} \max_{u \in \mathcal{S}^{d_1-1}, v \in \mathcal{S}^{d_2-1}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}|\xi_i Z_i| \leq C\nu_\delta \alpha^{-\delta}.
$$

Take $t = 2\log(7) \cdot (d_1 + d_2)$ in (43), and combined the above inequality gives

$$
\|\nabla L_{n,\alpha}(\Theta^*)\|_{op} \leq C\left(\sqrt{\frac{\nu_\delta \alpha^{1-\delta}(d_1+d_2)}{n}} + \frac{\alpha(d_1+d_2)}{n} + \nu_\delta \alpha^{-\delta}\right)
$$

with probability at least $1 - 2 \cdot 7^{-(d_1+d_2)}$. Thus, we complete the proof of(10). □

Proof of Theorem 2. Setting

$$
\begin{aligned}
\alpha &= C\left(\frac{n}{d_1+d_2}\right)^{1/(1+\min\{1,\delta\})}, \\
\lambda &= C\rho_u^{1/2}\left(\sqrt{\nu_\delta} + 1 + \nu_\delta\right)\alpha^{-\min\{1,\delta\}},
\end{aligned}
$$

in Proposition 3 yields

$$\|\nabla L_{n,\alpha}(\Theta^*)\|_{op} \leq \lambda/2$$

with probability at least $1 - 2 \cdot 7^{-(d_1+d_2)}$. Finally, applying the results in Theorem 1 completes the proof. □

References

[1] Olivier Bousquet, Concentration inequalities for sub-additive functions using the entropy method, in: Evariste Giné, Christian Houdré, David Nualart (Eds.), Stochastic Inequalities and Applications, Birkhäuser, Basel, 2003, pp. 213–247, ISBN 978-3-0348-8069-5.
[2] Olivier Catoni, Challenging the empirical mean and empirical variance: a deviation study, Ann. Inst. Henri Poincaré Probab. Stat. 48 (4) (2012) 1148–1185, https://doi.org/10.1214/11-AIHP454.
[3] Jianhui Chen, Jieping Ye, Sparse trace norm regularization, Comput. Stat. 29 (3–4) (2014) 623–639.
[4] Laming Chen, Yuantao Gu, The convergence guarantees of a non-convex approach for sparse recovery using regularized least squares, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 3350–3354.
[5] Y. Dodge, Least absolute deviation regression, The Concise Encyclopedia of Statistics (2008) 299–302.
[6] Jianqing Fan, Runze Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc. 96 (456) (2001) 1348–1360.
[7] Jianqing Fan, Quefeng Li, Yuyan Wang, Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions, J. R. Stat. Soc., Ser. B, Stat. Methodol. 79 (1) (2017) 247–265.
[8] Jianqing Fan, Han Liu, Qiang Sun, Tong Zhang, I-lamm for sparse learning: simultaneous control of algorithmic complexity and statistical error, Ann. Stat. 46 (2) (2018) 814.
[9] Jianqing Fan, Wenyan Gong, Ziwei Zhu, Generalized high-dimensional trace regression via nuclear norm regularization, J. Econom. 212 (1) (2019) 177–202.
[10] Yuwen Gu, Hui Zou, High-dimensional generalizations of asymmetric least squares regression and their applications, Ann. Stat. 44 (6) (2016) 2661–2694, https://doi.org/10.1214/15-AOS1431.
[11] Jianhua Guo, Jianchang Hu, Bing-Yi Jing, Zhen Zhang, Spline-lasso in high-dimensional linear regression, J. Am. Stat. Assoc. 111 (513) (2016) 288–297.
[12] Gui Huan, Jiawei Han, Quanquan Gu, Towards faster rates and oracle property for low-rank matrix estimation, in: Maria Florina Balcan, Kilian Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning, New York, New York, USA, 20–22 Jun 2016, in: Proceedings of Machine Learning Research, vol. 48, PMLR, 2016, pp. 2300–2309, https://proceedings.mlr.press/v48/gui16.html.
[13] Dongxiao Han, Jian Huang, Yuanyuan Lin, Guohao Shen, Robust post-selection inference of high-dimensional mean regression with heavy-tailed asymmetric or heteroskedastic errors, J. Econom. 230 (2) (2022) 416–431.
[14] Bruce M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Stat. (1975) 1163–1174.
[15] Jian Huang, Shuangge Ma, Hongzhe Li, Cun-Hui Zhang, The sparse Laplacian shrinkage estimator for high-dimensional regression, Ann. Stat. 39 (4) (2011) 2021.
[16] Peter J. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35 (1) (1964) 73–101, https://doi.org/10.1214/aoms/1177703732.
[17] Peter J. Huber, Robust regression: asymptotics, conjectures and Monte Carlo, Ann. Stat. 1 (5) (1973) 799–821, ISSN 00905364, http://www.jstor.org/stable/2958283.
[18] Roger W. Koenker, Gilbert W. Bassett Jr., Regression quantiles, Econometrica (1978) 33–50.
[19] Vladimir Koltchinskii, Karim Lounici, Alexandre B. Tsybakov, Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, Ann. Stat. 39 (5) (2011) 2302–2329, https://doi.org/10.1214/11-AOS894.
[20] Michel Ledoux, Michel Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, vol. 23, Springer Science & Business Media, Berlin, Heidelberg, 1991.
[21] Xin Li, Dongya Wu, Chong Li, Jinhua Wang, Jen-Chih Yao, Sparse recovery via nonconvex regularized m-estimators over l_q-balls, Comput. Stat. Data Anal. 152 (2020) 107047.
[22] Po-Ling Loh, Scale calibration for high-dimensional robust regression, Electron. J. Stat. 15 (2) (2021) 5933–5994.
[23] Po-Ling Loh, Martin J. Wainwright, Regularized m-estimators with nonconvexity: statistical and algorithmic theory for local optima, Adv. Neural Inf. Process. Syst. 26 (2013).
[24] Po-Ling Loh, Martin J. Wainwright, Support recovery without incoherence: a case for nonconvex regularization, Ann. Stat. 45 (6) (2017) 2455–2482, https://doi.org/10.1214/16-AOS1530.
[25] Rebeka Man, Kean Ming Tan, Zian Wang, Wen-Xin Zhou, Retire: robust expectile regression in high dimensions, J. Econom. (2023) 105459, https://doi.org/10.1016/j.jeconom.2023.04.004, ISSN 0304-4076.
[26] Shike Mei, Bin Cao, Jiantao Sun, Encoding low-rank and sparse structures simultaneously in multi-task learning, Adv. Neural Inf. Process. Syst. (2012).
[27] Sahand Negahban, Martin J. Wainwright, Restricted strong convexity and weighted matrix completion: optimal bounds with noise, J. Mach. Learn. Res. 13 (1) (2012) 1665–1697.
[28] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, Bin Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, Stat. Sci. 27 (4) (2012) 538–557, https://doi.org/10.1214/12-STS400.

[29] Ling Peng, Xiangyong Tan, Peiwen Xiao, Zeinab Rizk, Xiaohui Liu, Oracle inequality for sparse trace regression models with exponential β-mixing errors, Acta Math. Sin. Engl. Ser. 39 (2023) 2031–2053.
[30] Qiang Sun, Wen-Xin Zhou, Jianqing Fan, Adaptive Huber regression, J. Am. Stat. Assoc. 115 (529) (2020) 254–265.
[31] Xiangyong Tan, Ling Peng, Peiwen Xiao, Qing Liu, Xiaohui Liu, The rate of convergence for sparse and low-rank quantile trace regression, J. Complex. 79 (2023) 101778.
[32] A.W. van der Vaart, Jon A. Wellner, Weak Convergence and Empirical Processes: With Applications to Statistics, vol. 3, Springer, 1996.
[33] Lili Wang, Chao Zheng, Wen Zhou, Wen-Xin Zhou, A new principle for tuning-free Huber regression, Stat. Sin. 31 (4) (2021) 2153–2177.
[34] Yaohong Yang, Weihua Zhao, Lei Wang, Online regularized matrix regression with streaming data, Comput. Stat. Data Anal. 187 (2023) 107809, https://doi.org/10.1016/j.csda.2023.107809, ISSN 0167-9473.
[35] Cun-Hui Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Stat. 38 (2) (2010) 894–942, https://doi.org/10.1214/09-AOS729.
[36] Tong Zhang, Analysis of multi-stage convex relaxation for sparse regularization, J. Mach. Learn. Res. 11 (35) (2010) 1081–1107, http://jmlr.org/papers/v11/zhang10a.html.
[37] Junlong Zhao, Lu Niu, Shushi Zhan, Trace regression model with simultaneously low rank and row (column) sparse parameter, Comput. Stat. Data Anal. 116 (2017) 1–18.
[38] Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, Han Liu, A new perspective on robust m-estimation: finite sample theory and applications to dependence-adjusted multiple testing, Ann. Stat. 46 (5) (2018) 1904.