Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 18, No. 2, pp. 439–458, 2022

LOW-RANK TENSOR HUBER REGRESSION*

Yangxin Wei, Ziyan Luo† and Yang Chen

Abstract: Low-rank tensor regression has been well considered under general least squares framework, but it is highly sensitive to the outliers or heavy-tailed errors. To tackle this problem, we propose a low-rank tensor Huber regression model in which the tensor nuclear norm regularization is employed to characterize the low-rankness. The risk bound of the resulting estimator is established under mild assumptions. In addition, an efficient and stable alternating direction method of multipliers based algorithm is designed to solve the proposed model, and the global convergence as well as the computational complexity of the algorithm is also analyzed. Finally, numerical experiments conducted both on synthetic data with different types of noises and a real dataset illustrate the robustness and effectiveness of the approach. Especially when the noise is heavy-tailed or the coefficient tensor is low-rank, the mean square error of the estimator obtained by our model can be orders of magnitude better than several existing methods.

Key words: *tensor Huber regression, low-rank, heavy-tailed noise, risk bound, alternating direction method of multipliers*

Mathematics Subject Classification: *15A18, 15A69, 65F15, 90C33*

1 Introduction

To solve the regression problem with tensor data, researchers have proposed tensor regression, which widely exists in scientific research and practical application fields, such as recommendation system [4], medical imaging data analysis [17], image/video analysis [29], machine learning and artificial intelligence [34]. The ordinary regularized tensor least squares, such as [11, 17, 36, 41], are highly sensitive to the outliers or heavy-tailed errors. In this sense, it is natural to consider the robust approach, such as the popular quantile regression [21] and least absolute deviation regression [20]. Meanwhile, by imposing sparsity and/or low rank structures on parameters, low dimensional tensor regression problem has been widely studied in the development of algorithms and theoretical guarantee in recent years, see Buhlmann and van de Geer [5] for an overview. However, it is much less clear how to effectively handle the scenario when outliers/heavy-tailed errors and low dimensional structures are simultaneously inherent to these regression problems. The main purpose of this paper is to provide a robust method for low-rank coefficient tensor regression.

Comparing with the quantile regression and least absolute deviation regression, regression with Huber loss [18, 19, 38] has received more and more attention due to its adaptability

*This work was supported by the National Natural Science Foundation of China (No. 11771038) and Beijing Natural Science Foundation (No. Z190002).

†Corresponding author

to errors with different shapes and tails. Related research includes the use of Huber loss together with the adaptive lasso penalty [25], the regularized approximate quadratic estimator with an ℓ_1-penalty (RA Lasso) [10], ℓ_1-norm regularized vector Huber regression model with linear equality and inequality constraints [30], and ℓ_0-norm regularized vector Huber regression model [1]. Among them, ℓ_1- or ℓ_0-penalty is used to control the inherited sparsity of the estimator, thereby alleviating the over-fitting issue and/or achieving data dimension reduction. Moreover, there is now a substantial body of work on low-rank matrix Huber regression, in which the nuclear norm regularization is often adopted to control the low-rank structure of the coefficient matrix. For instance, Elsener and van de Geer [9] have studied the nuclear norm regularized single response Huber regression model and proved the sharp Oracle inequality of risk function, Chen et al. [7] have proposed a low-rank elastic-net regularized multivariate Huber regression model and designed an accelerated proximal gradient algorithm.

To sum up, little is known about statistical theory as well as efficient algorithms for estimation of the low-rank tensor Huber regression models, although vector and matrix methods have been extensively studied. However, those works on vector or matrix models cannot be extended to tackle tensor models directly. On the one hand, the method of unfolding tensor into vector or matrix will break down the spatial structure, and result in the loss of information. On the other hand, it will lead to the dimensionality disaster and then the over-fitting phenomenon, especially for small sample size data [31, 44]. This motivates us to establish robust low-rank tensor Huber regression for original tensor data which is friendly to outliers/heavy tailed errors.

Focusing on the low-rank tensor regression, numerous researchers have adopted nuclear norm regularization techniques to enhance the low-rank structure, see, e.g., [23, 28, 29, 37]. It is worth pointing out that the existing statistical properties and numerical algorithms along this line are mostly concentrated on ordinary least squares. For example, Raskutti et al. [36] have deduced the statistical upper bound for the low-rank least squares tensor model with the nuclear norm penalty, and Li et al. [26] have investigated the estimation error upper bound for the proposed tensor response linear model with the nuclear-ℓ_1-norm regularization and developed an M-ADMM-based algorithm to achieve low-rank and sparse tensor recovery. Little work has addressed the low-rank robust tensor regression.

This motivates us to build the nuclear norm regularized tensor Huber regression (NNTH) model for robust estimation. The resulting NNTH estimator will be shown to possess nice risk bound theoretically. In addition, to compute NNTH estimator, an efficient tensor alternating direction method of multipliers (ADMM) algorithm is designed. The contributions of this paper include: (1) An NNTH model is proposed to deal with original tensor data directly. This model not only preserves low-rank structure of tensor data, but also reduces the negative impact of outliers. (2) The risk bound of the resulting estimator is established. (3) An ADMM algorithm is designed which enjoys low computational complexity and global convergence.

The remainder of the paper is organized as follows. In Section 2, the NNTH model is introduced for the low-rank tensor regression with outliers/heavy-tailed errors. In Section 3, we establish the risk bound for the resulting NNTH estimator. In Section 4, we design an ADMM-based algorithm to solve the proposed NNTH model, and analyze the convergence as well as the computational complexity of the algorithm. Simulation studies and a real data analysis are discussed in Section 5. A brief conclusion is drawn in Section 6. For convenience, notation that will be used throughout the paper is listed in Table 1.

Table 1: A list of notation.

$:=$	Defined as
$\mathbb{R}^m$	The m-dimensional real vector space
$\mathbb{R}^{m\times q}$	The $m \times q$-dimensional real matrix space
$\mathbb{R}^{I_1\times\cdots\times I_M}$	The $I_1 \times \cdots \times I_M$-dimensional real tensor space
$\mathbb{I}_d$	The identity matrix of dimension d
$\bar{\mathbb{M}}$	The closure of subspace $\mathbb{M}$
$\mathbb{M}^\perp$	The orthogonal complement of subspace $\mathbb{M}$
$\langle\cdot,\cdot\rangle$	The inner product
$\otimes$	The Kronecker product
$\lVert\cdot\rVert_*$	The matrix/tensor nuclear norm
$\lVert\cdot\rVert$	The matrix spectral norm
$\lVert\cdot\rVert_*^*$	The dual norm of tensor $\lVert\cdot\rVert_*$
$\lVert\cdot\rVert_F$	The Frobenius norm
$\lVert\cdot\rVert_2$	The Euclidean norm of vectors
$\nabla f(\cdot)$	The gradient of function f
$\partial f(\cdot)$	The subdifferential of function f
$\nabla^2 f(\cdot)$	The Hessian matrix of function f
$Prox_{\beta f}(\cdot)$	The proximal operator of f associated with a parameter β
$[n]$	The index set $\{1,\ldots,n\}$
$\mathtt{vec}(X)$	The column vector generated by stacking all columns of the matrix X
$vtt(\cdot)$	The inverse transformation of $\mathtt{vec}(\cdot)$
$\mathbb{E}(x)$	The expectation of random variable x
A^T	The transpose of the matrix A
$B_{(d)}$	The d-mode unfolding of the tensor $\mathcal{B}$
$fold_d(B_{(d)})$	The inverse operation of $B_{(d)}$
$\mathtt{rank}(\mathcal{B})$	The Tucker rank of tensor $\mathcal{B}$
$\sigma(A)$	The singular value of the matrix A
$\lambda_{min}(M)$	The minimum eigenvalue of square matrix M
$\lambda_{max}(M)$	The maximum eigenvalue of square matrix M
$Pr(A)$	The probability of the occurrence of the event A
$Pr(A\vert B)$	The probability of the event A under the precondition B
$\lvert\mathcal{A}\rvert$	The number of elements in set $\mathcal{A}$
$\mathcal{A}\times_d B$	The d-mode product of the tensor $\mathcal{A}$ and the matrix B

2 Methodology

2.1 Tensor Basics

An Mth-order tensor $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_M}$ is an M-way array consisting of entries $x_{i_1\ldots i_M}$ with each i_j varying among $1,\ldots,I_j$ for all $j \in [M]$. Vectors and matrices are typical low order tensors with $M = 1$ and $M = 2$, respectively. Some useful operations that transform a tensor into a matrix or a vector are recalled. The operator $\mathtt{vec}(\mathcal{X})$ stacks the entries of $\mathcal{X}$ into a $\prod_m I_m$ dimensional column vector. The m-mode unfolding, termed as $X_{(m)}$, maps a tensor $\mathcal{X}$ into a $I_m \times \prod_{m'\neq m} I_{m'}$ matrix. Define the inner product and the Frobenius norm for tensors with symbols $\langle\cdot,\cdot\rangle$, $\|\cdot\|_F$, respectively, formulated by

$$\langle\mathcal{X},\mathcal{Y}\rangle := \sum_{i_1=1}^{I_1}\cdots\sum_{i_M=1}^{I_M} x_{i_1\ldots i_M} y_{i_1\ldots i_M} \in \mathbb{R}, \quad \|\mathcal{X}\|_F := \sqrt{\langle\mathcal{X},\mathcal{X}\rangle},$$

for all $\mathcal{X},\mathcal{Y} \in \mathbb{R}^{I_1\times\cdots\times I_M}$. The m-mode product of the tensor $\mathcal{X}$ with a matrix $U \in \mathbb{R}^{R_m\times I_m}$, termed as $\mathcal{X}\times_m U$, yields a tensor $\mathcal{Y} \in \mathbb{R}^{I_1\times I_2\times\cdots\times R_m\times\cdots\times I_M}$ with its entries $y_{i_1 i_2\ldots r_m\ldots i_M} = \sum_{i_m=1}^{I_m} \mathcal{X}_{i_1 i_2\ldots i_m\ldots i_M} U_{r_m i_m}$.

The Tucker decomposition is one of the most important decompositions for general high order tensors. For a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_M}$, there exists a tensor $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_M}$ and M matrices $U_m \in \mathbb{R}^{I_m \times r_m} (m \in [M])$ (usually have orthogonal columns, i.e., $U_m^T U_m = \mathbb{I}_{r_m}, m \in [M]$), such that

$$\mathcal{X} = \mathcal{G} \times_1 U_1 \times_2 \cdots \times_M U_M.$$

Introducing the Tucker rank of $\mathcal{X}$, written as $\mathtt{rank}(\mathcal{X})$, with definitional expression as the vector $(\mathtt{rank}(X_{(1)}), \ldots, \mathtt{rank}(X_{(M)}))$, where each $\mathtt{rank}(X_{(m)})$ is called the m-rank of $\mathcal{X}$. Recall from [29] that the tensor nuclear norm is defined as

$$\|\mathcal{X}\|_* := \frac{1}{M} \sum_{m=1}^{M} \|X_{(m)}\|_*. \tag{2.1}$$

Declared by [42, Lemma 1], the dual norm of the nuclear norm is defined as

$$\|\mathcal{X}\|_*^* := \inf_{\frac{1}{M}\left(\mathcal{Y}^{(1)} + \cdots + \mathcal{Y}^{(M)}\right) = \mathcal{X}} \max_{d=1,\ldots,M} \left\| Y_{(d)}^{(d)} \right\|, \tag{2.2}$$

where $Y_{(d)}^{(d)}$ is the d-mode unfolding of $\mathcal{Y}^{(d)}$. Moreover, it has been shown that

$$\|\mathcal{X}\|_*^* \leq \frac{1}{M} \sum_{d=1}^{M} \|X_{(d)}\| \leq \max_{d=1,\ldots,M} \|X_{(d)}\|. \tag{2.3}$$

More tensor basics can be found in [22, 35].

2.2 Low Rank Regularized Tensor Huber Regression

Consider the tensor regression problem in which covariate tensors $\mathcal{X}_i \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ and responses $y_i \in \mathbb{R}$ are related by

$$y_i = \langle \mathcal{B}, \mathcal{X}_i \rangle + \epsilon_i, \quad \forall i \in [N], \tag{2.4}$$

where $\{\mathcal{X}_i : i \in [N]\}$ are independent and identically distributed (*i.i.d.*) covariate tensors, $\{\epsilon_i : i \in [N]\}$ are *i.i.d.* errors, and $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ is the unknown coefficient tensor. The distributions of the random covariate tensor $\mathcal{X}$ and the random variable $\epsilon|\mathcal{X}$ (the random error conditioning on $\mathcal{X}$) are both assumed to have mean zero.

Outliers are frequently encountered in practical problems, as claimed by Hampel et al. [14] that a general dataset contains about $1\% - 10\%$ or more outliers. Here, outliers refer to the points inconsistent with the general behavior or characteristics of other points in the sample space $\{y_i : i \in [N]\}$ due to the external interference (see, e.g.,[14, 2]). Besides the influence of outliers, the random error ϵ may be heavy-tailed, which means that the moment generating function $\mathrm{E}\{\exp(t\epsilon)\} = \infty$ for all $t > 0$ [13]. Typical heavy-tailed distributions include the LogNormal distribution, and the Weibull distribution with shape parameter in $(0, 1)$. Taking outliers and heavy-tailed errors into consideration, it is natural to adopt the Huber loss function to estimate the coefficient tensor $\mathcal{B}$ based on observations $\{(\mathcal{X}_i, y_i) : i \in [N]\}$. The resulting optimization model is

$$\min_{\mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_M}} \frac{1}{N} \sum_{i=1}^{N} h_\alpha(y_i - \langle \mathcal{B}, \mathcal{X}_i \rangle),$$

where h_α is the Huber function defined as

$$h_\alpha(z) = \begin{cases} \frac{1}{2}|z|^2, & \text{if } |z| \leq \alpha, \\ \alpha(|z| - \frac{1}{2}\alpha), & \text{otherwise.} \end{cases} \tag{2.5}$$

Here, $\alpha > 0$ is the robustification parameter that controls the blending of the quadratic loss (bias) and the absolute loss (robustness) [18, 19]. To better illustrate the effect of the Huber loss, the proximal operator of h_α is analyzed. Recall from [3, Definition 12.23] that the proximal operator of a function $f : \mathbb{R} \to \mathbb{R}$, associated with a parameter $\beta > 0$, at point $x \in \mathbb{R}$, is defined by

$$\texttt{Prox}_{\beta f}(x) = \arg\min_{y \in \mathbb{R}} \left\{ \beta f(y) + \frac{1}{2}(y - x)^2 \right\}. \tag{2.6}$$

Direct manipulations lead to the following explicit formula of $\texttt{Prox}_{\beta h_\alpha}$.

Lemma 2.1. *Given α, $\beta > 0$, and $x \in \mathbb{R}$, we have*

$$\texttt{Prox}_{\beta h_\alpha}(x) = \begin{cases} x - \beta\alpha, & \textit{if } x > (1+\beta)\alpha, \\ \frac{x}{\beta+1}, & \textit{if } |x| \leq (1+\beta)\alpha, \\ x + \beta\alpha, & \textit{otherwise.} \end{cases}$$

The graph of $\texttt{Prox}_{\beta h_\alpha}$ is plotted in Figure 1 with various α at $\beta = 3$. As we can see, the blue solid line($\alpha = 0$) refers to the case of the least squares loss, and the effect of graph shrinkage is intensified with the increase of α.

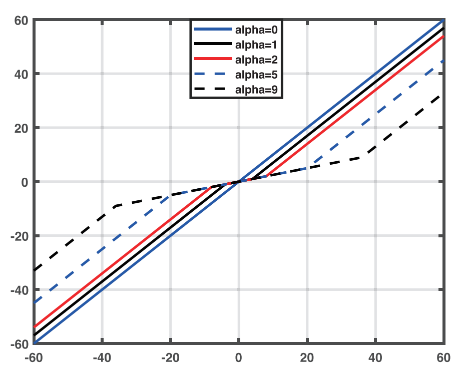


Figure 1: $\texttt{Prox}_{\beta h_\alpha}(x)$ at $\beta = 3$.

It is worth mentioning that in the foregoing tensor regression model, the number of parameters $\Pi_{d=1}^M I_d$ is often larger than the sample size N, and the resulting high-dimensional setting inspires us to incorporate "sparsity" to further reduce the number of parameters of interest. Here, the low-rankness based on tensor Tucker decomposition (see, e.g., [22]) is adopted for sparsity characterization of the coefficient tensor $\mathcal{B}$. The reasons for using low-Tucker-rankness are three-fold. The first one is the tractability of Tucker decomposition, using singular value decomposition on all unfolding matrices. The second one is the flexibility of low-rankness, allowing different values of ranks along different modes. The third one is the applicability of practical datasets, for instance, the application in neuroimaging analysis

[27]. In such senses, the estimator of the low-rank regularized Huber regression can be solved by

$$\min_{\mathcal{B}\in\mathbb{R}^{I_1\times\cdots\times I_M}} \frac{1}{N}\sum_{i=1}^{N} h_\alpha(y_i - \langle \mathcal{B}, \mathcal{X}_i\rangle) + \lambda\|\mathcal{B}\|_*. \tag{2.7}$$

Here $\|\mathcal{B}\|_*$ is the tensor nuclear norm of $\mathcal{B}$ which serves as a convex surrogate of the Tucker rank, and $\lambda > 0$ is the regularization parameter. Thus, the resulting approach is termed as nuclear norm regularized tensor Huber regression (NNTH for short).

3 Risk Bounds

This section is devoted to the risk bound analysis of the NNTH estimator generated by the optimal solution to problem (2.7). For convenience, we denote $y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N$, $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_N) \in \mathbb{R}^{I_1\times\cdots\times I_M\times N}$. Denote the true coefficient tensor by $\mathcal{B}^*$, and the NNTH estimator simply by $\hat{\mathcal{B}}$. Recall that the Huber regression coefficient tensor is given by

$$\mathcal{B}^*_\alpha \in \mathop{\arg\min}_{\mathcal{B}\in\mathbb{R}^{I_1\times\cdots\times I_M}} \mathbb{E}\left\{h_\alpha(y - \langle \mathcal{B}, \mathcal{X}\rangle)\right\},$$

where the expectation is taken over the regression errors. The statistical error $\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F$ is then bounded by

$$\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F \le \|\hat{\mathcal{B}} - \mathcal{B}^*_\alpha\|_F + \|\mathcal{B}^*_\alpha - \mathcal{B}^*\|_F, \tag{3.1}$$

where the first term on the right-hand side is the estimation error, and the other term is the approximation error. In what follows, upper bounds of these two errors will be given so as to arrive the risk bound of $\hat{\mathcal{B}}$. To proceed, some moment conditions on $\mathcal{X}$ and $\epsilon|\mathcal{X}$ are introduced as below, which are adopted from [10].

Condition 3.1. $\mathbb{E}\{\mathbb{E}(|\epsilon|^k|\mathcal{X})\}^2 \le M_k < \infty$, for some $k \ge 2$.

Condition 3.2. $0 < \kappa_l = \lambda_{min}(\mathbb{E}(A)) \le \lambda_{max}(\mathbb{E}(A)) = \kappa_u < \infty$ with $A = \mathtt{vec}(\mathcal{X})\mathtt{vec}(\mathcal{X})^T$.

Condition 3.3. For any $\mathcal{V} \in \mathbb{R}^{I_1\times\cdots\times I_M}$, $\langle \mathcal{X}, \mathcal{V}\rangle$ is sub-Gaussian with parameter at most $\kappa_0^2\|\mathcal{V}\|_F^2$, i.e., $\mathbb{E}\{\exp(t\langle \mathcal{X}, \mathcal{V}\rangle)\} \le \exp(t^2\kappa_0^2\|\mathcal{V}\|_F^2/2)$, for any $t \in \mathbb{R}$.

It is worth pointing out that Condition 3.1 is valid for most common distributions, such as the normal distribution, the Weibull distribution and the LogNormal distribution. Approximation error can be inferred from [10] by taking $\beta^*_\alpha = \mathtt{vec}(\mathcal{B}^*_\alpha)$ and $\beta^* = \mathtt{vec}(\mathcal{B}^*)$ as follows.

Theorem 3.4. *Under Conditions 3.1–3.3 there is an absolute positive constant C_1, such that*

$$\|\mathcal{B}^*_\alpha - \mathcal{B}^*\|_F \le C_1\sqrt{\kappa_u}\kappa_l^{-1}(\kappa_0^k + \sqrt{M_k})\alpha^{1-k}, \tag{3.2}$$

where k, κ_u, κ_l, κ_0, M_k are defined as in Conditions 3.1–3.3.

According to the expression on the right side of inequality in Theorem 3.4, when $\alpha \to \infty$, the error $\|\mathcal{B}^*_\alpha - \mathcal{B}^*\|_F \to 0$, that is, $\mathcal{B}^*_\alpha \to \mathcal{B}^*$. Furthermore, this theorem also shows that if the higher-order moment of error exists, the approximation error will decrease rapidly when α increases.

As declared by Negahban et al. [32], the decomposability of the regularizer and the restricted strong convexity (RSC) are two key properties for establishing a sharp convergence result for a regularized M-estimator. Before embarking on the upper bound of the estimation error $\|\hat{\mathcal{B}} - \mathcal{B}^*_\alpha\|_F$, the d-mode decomposability in [42] is recalled as below.

Denote $\hat{\triangle} := \hat{\mathcal{B}} - \mathcal{B}^*_\alpha \in \mathbb{R}^{I_1 \times \cdots \times I_M}$. For each $d \in [M]$, let $B^*_{\alpha(d)} = U_d S_d V_d^T$ be the condensed singular value decomposition of $B^*_{\alpha(d)}$ with $U_d \in \mathbb{R}^{I_d \times r_d}$ and $V_d \in \mathbb{R}^{(\prod_{d' \neq d} I_{d'}) \times r_d}$, where r_d is the rank of $B^*_{\alpha(d)}$. Set

$$\hat{\triangle}''_d = \left(\mathbb{I}_{I_d} - U_d U_d^T\right) \hat{\triangle}_{(d)} \left(\mathbb{I}_{\prod_{d' \neq d} I_{d'}} - V_d V_d^T\right), \text{ and } \hat{\triangle}'_d := \hat{\triangle}_{(d)} - \hat{\triangle}''_d, \tag{3.3}$$

then

$$\|B^*_{\alpha(d)} + \hat{\triangle}''_d\|_* = \|B^*_{\alpha(d)}\|_* + \|\hat{\triangle}''_d\|_*. \tag{3.4}$$

Together with the triangle inequality with respect to $\|\cdot\|_*$, we have

$$\|\hat{\mathcal{B}}\|_* = \frac{1}{M}\sum_{d=1}^{M} \|\hat{\triangle}'_d + \hat{\triangle}''_d + B^*_{\alpha(d)}\|_* \geq \frac{1}{M}\sum_{d=1}^{M}\left(-\|\hat{\triangle}'_d\|_* + \|\hat{\triangle}''_d\|_* + \|B^*_{\alpha(d)}\|_*\right). \tag{3.5}$$

Meanwhile, similar to [42, Lemma 2] for nuclear norm regularized tensor least squares estimator, the NNTH estimator $\hat{\mathcal{B}}$ also possesses the following property.

Lemma 3.5. *If $\lambda \geq 2\|\nabla H_\alpha(\mathcal{B}^*_\alpha)\|^*_*$, then $\mathtt{rank}\left(\hat{\triangle}'_d\right) \leq 2\ \mathtt{rank}\left(B^*_{\alpha(d)}\right)$ for each $d \in [M]$, and $\sum_{d=1}^M \|\hat{\triangle}''_d\|_* \leq 3\ \sum_{d=1}^M \|\hat{\triangle}'_d\|_*$.*

Proof. By mimicking the proof of [33, Lemma 1 (a)], we can obtain that for any $d \in [M]$, $\mathtt{rank}\left(\hat{\triangle}'_{(d)}\right) \leq 2\ \mathtt{rank}\left(B^*_{\alpha(d)}\right)$, which indicates $\mathtt{rank}(\hat{\triangle}') \leq 2\ \mathtt{rank}(\mathcal{B}^*_\alpha)$. To derive the second part, one has

$$\begin{aligned} -\|\nabla H_\alpha(\mathcal{B}^*_\alpha)\|^*_* \cdot \|\hat{\triangle}\|_* &\leq \left\langle \nabla H_\alpha(\mathcal{B}^*_\alpha), \hat{\triangle}\right\rangle \leq H_\alpha\left(\hat{\triangle} + \mathcal{B}^*_\alpha\right) - H_\alpha\left(\mathcal{B}^*_\alpha\right) \\ &\leq \lambda\left(\|\mathcal{B}^*_\alpha\|_* - \|\hat{\mathcal{B}}\|_*\right) \leq \frac{\lambda}{M}\sum_{d=1}^{M}\left(\|\hat{\triangle}'_d\|_* - \|\hat{\triangle}''_d\|_*\right) \end{aligned}$$

where the first inequality is from the Hölder inequality deduced from the dual norm, the second inequality is from convexity of the Huber loss function, the third inequality is from the optimality of $\hat{\mathcal{B}}$ to problem (2.7), and the last one is from (3.5). Along with $\lambda \geq 2\|\nabla H_\alpha(\mathcal{B}^*_\alpha)\|^*_*$, the desired inequality follows readily from the triangle inequality $\|\hat{\triangle}_{(d)}\|_* \leq \|\hat{\triangle}'_d\|_* + \|\hat{\triangle}''_d\|_*$. This completes the proof. ☐

Define the following two constraint sets regarding to $\triangle$,

$$\mathscr{C}(\triangle) := \{\triangle \in \mathbb{R}^{I_1 \times \cdots \times I_M} : \|\triangle''\|_* \leq 3\|\triangle'\|_*\}, \text{ and } \mathscr{B}(\triangle) := \{\triangle \in \mathbb{R}^{I_1 \times \cdots \times I_M} : \|\triangle\|_F \leq 1\}. \tag{3.6}$$

Fan et al. [10] have proved that the vector Huber loss function satisfies RSC under some conditions by using [32, Lemma 2]. This property can be extended to the tensor case.

Lemma 3.6. *Suppose that Conditions 3.1 – 3.3 hold, then the RSC condition*

$$\delta H_\alpha(\triangle, \mathcal{B}) := H_\alpha(\mathcal{B} + \triangle) - H_\alpha(\mathcal{B}) - \langle \nabla H_\alpha(\mathcal{B}), \triangle\rangle \geq \kappa_H \|\triangle\|_F^2 - \tau_H \|\triangle\|_*^2 \tag{3.7}$$

holds for any tensor $\triangle \in \mathscr{C}(\triangle) \cap \mathscr{B}(\triangle)$ with $\kappa_H = \frac{\kappa_1}{4}, \tau_H = 8\kappa_2 \frac{\log\left(\sum_{d=1}^M I_d\right)}{N}$, where $\kappa_1, \kappa_2, c'_1, c'_2$ are absolute positive constants.

Lemma 3.6 shows that RSC holds with absolute constants that do not depend on α, which makes a key role in the following theorem. Combining with the decomposability of the nuclear norm and RSC of the Huber loss function, we can give an upper bound of the estimation error as below.

Theorem 3.7. *Under the conditions used in Lemmas* 3.5 − 3.6 *and* $\|\mathcal{B}\|_* \le R_1$*, there are absolute positive constants* C_2 *and* C_3 *such that*

$$\|\hat{\mathcal{B}} - \mathcal{B}_\alpha^*\|_F \le \left(C_2\kappa_2 R_1 N^{-1} log\left(\sum_{d=1}^M I_d\right) + C_3\lambda \right) \kappa_1^{-1} M^{-1} \sum_{d=1}^M \sqrt{r_d}. \tag{3.8}$$

Proof. Define a function $F : \mathbb{R}^{I_1\times\cdots\times I_M} \to \mathbb{R}$,

$$F(\triangle) = H_\alpha(\mathcal{B}_\alpha^* + \triangle) - H_\alpha(\mathcal{B}_\alpha^*) + \lambda(\|\mathcal{B}_\alpha^* + \triangle\|_* - \|\mathcal{B}_\alpha^*\|_*) \tag{3.9}$$

Because of $F(0) = 0$, the error $\hat{\triangle} = \hat{\mathcal{B}} - \mathcal{B}_\alpha^*$ satisfies $F(\hat{\triangle}) \le F(0) = 0$. Due to Lemma 3.6, Cauchy-Schwarz inequality, triangle inequality, $\lambda \ge 2\|\nabla H_\alpha(\mathcal{B}_\alpha^*)\|_*^*$ and $\|\mathcal{B}\|_* \le R_1$,

$$\begin{aligned} F(\hat{\triangle}) &\ge \left\langle \nabla H_\alpha(\mathcal{B}_\alpha^*), \hat{\triangle} \right\rangle + \kappa_H\|\hat{\triangle}\|_F^2 - \tau_H\|\hat{\triangle}\|_*^2 - \lambda\|\hat{\triangle}\|_* \\ &\ge \kappa_H\|\hat{\triangle}\|_F^2 - \left(2\tau_H R_1 + \frac{3\lambda}{2}\right)\|\hat{\triangle}\|_*. \end{aligned} \tag{3.10}$$

From Lemma 3.5, we can get an upper bound of $\|\hat{\triangle}\|_*$ by

$$\|\hat{\triangle}\|_* \le 4\|\hat{\triangle}'\|_* \le \frac{4}{M}\sum_{d=1}^M \sqrt{2r_d}\|\hat{\triangle}'_{(d)}\|_F \le \frac{4\|\hat{\triangle}\|_F}{M}\sum_{d=1}^M \sqrt{2r_d}.$$

Taking $\kappa_H = \frac{\kappa_1}{4}$ and $\tau_H = 8\kappa_2\frac{\log(\sum_{d=1}^M I_d)}{N}$, we obtain the desired assertion by (3.9) and (3.10). □

The following lemma will serve a more reasonable value of λ comparing to that in Lemma 3.5.

Lemma 3.8. *If Condition* 3.3 *holds, then there are absolute constants* $C', c_1, c_2 > 0$ *for a sample size* N *makes*

$$Pr\left(C'\alpha\kappa_0\sqrt{\frac{I_j + \prod_{j'\ne j} I_{j'}}{N}} \ge 2\|\nabla H_\alpha(\mathcal{B}_\alpha^*)\|_*^* \right) \ge 1 - c_1\exp\left\{ -c_2\left(\frac{I_j + \prod_{j'\ne j} I_{j'}}{N}\right)\right\}, \tag{3.11}$$

where $j := \underset{d\in[M]}{\arg\min}\|X_{(d)}\|$*, and* $I_j + \prod_{j'\ne j} I_{j'}$ *is the sum of rows and columns of* $\|X_{(j)}\|$*.*

Proof. By invoking $\nabla H_\alpha(\mathcal{B}) = -\frac{1}{N}\sum_{i=1}^N h'_\alpha(y_i - \langle\mathcal{B}, \mathcal{X}_i\rangle)\cdot\mathcal{X}_i$, (2.2) and (2.3), we can get

$$\nabla\|H_\alpha(\mathcal{B}_\alpha^*)\|_*^* \le \frac{1}{N}\sum_{i=1}^N \|X_{i(j)}\cdot h'_\alpha(y_i - \langle\mathcal{B}_\alpha^*, \mathcal{X}_i\rangle)\|, \tag{3.12}$$

where $X_{i(j)}$ is the j-mode unfolding of $\mathcal{X}_i$. Analog to the proof of [33, Lemma 3], let $S^{m-1} := \{u \in \mathbb{R}^m | \|u\|_2 = 1\}$ denote the unit Euclidean sphere. The norm has the variational representation

$$\begin{aligned}\|X_{i(j)} \cdot h'_\alpha(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle)\| &= \sup_{u \in S^{I_j - 1}, v \in S^{\prod_{j' \neq j} I_{j'} - 1}} u_i^T X_{i(j)}^T h'_\alpha(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle) v_i \\ &\leq \sup_{u \in S^{I_j - 1}, v \in S^{\prod_{j' \neq j} I_{j'} - 1}} |\langle X_{i(j)} u_i, h'_\alpha(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle) v_i\rangle|. \quad (3.13)\end{aligned}$$

Let $\mathcal{N}_1$ and $\mathcal{N}_2$ denote 1/4 coverings of $S^{I_j - 1}$ and $S^{\prod_{j' \neq j} I_{j'} - 1}$, respectively. The coverage numbers are $|\mathcal{N}_1| \leq \left(1 + \frac{2}{1/4}\right)^{I_j} = 9^{I_j}$, $|\mathcal{N}_2| \leq \left(1 + \frac{2}{1/4}\right)^{\prod_{j' \neq j} I_{j'}} = 9^{\prod_{j' \neq j} I_{j'}}$. We now claim that

$$\nabla \|H_\alpha(\mathcal{B}^*_\alpha)\|_*^* \leq \frac{1}{N} \sum_{i=1}^N \sup_{u \in \mathcal{N}_1, v \in \mathcal{N}_2} |\langle X_{i(j)} u_i, h'_\alpha(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle) v_i\rangle|. \quad (3.14)$$

Select $u \in \mathcal{N}_1 \subseteq S^{I_j - 1}$ and $v \in \mathcal{N}_2 \subseteq S^{\prod_{j' \neq j} I_{j'} - 1}$. It can be calculated that

$$\frac{1}{\prod_{j' \neq j} I_{j'}} \left\langle X_{i(j)} u_i, h'_\alpha\left(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle\right) v_i \right\rangle = \frac{1}{\prod_{j' \neq j} I_{j'}} \sum_{k=1}^{\prod_{j' \neq j} I_{j'}} \left(x_{i(j)k}^T u_i\right) \left(v_i^T h'_\alpha\left(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle\right)\right)$$

and for any $k = 1, \ldots, \prod_{j' \neq j} I_{j'}$,

$$\mathbb{E}\left\{\left(x_{i(j)k}^T u_i\right)\left(v_i^T h'_\alpha\left(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle\right)\right)\right\} = 0, \quad \left|v_i^T h'_\alpha\left(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle\right)\right| \leq \alpha.$$

It is known from [43, Theorem 2.6.3] that, for any $t > 0$, there exists an absolute constant $C_1 > 0$ such that

$$Pr\left(\left|\frac{1}{\prod_{j' \neq j} I_{j'}} \langle X_{i(j)} u_i, h'_\alpha\left(y_i - \langle \mathcal{B}^*_\alpha, \mathcal{X}_i\rangle\right) v_i\rangle\right| \geq t \middle| X_{i(j)}\right) \leq 2\exp\left\{-\frac{C_1 \left(\prod_{j' \neq j} I_{j'}\right)^2 t^2}{\alpha^2 \|X_{i(j)} u_i\|_2^2}\right\}. \quad (3.15)$$

From Condition (3.3) and [43, Corollary 2.8.3], for any $s > 0$, there is an absolute constant $C_2 > 0$ which satisfies

$$\begin{aligned}&Pr\left(\left|\frac{1}{\prod_{j' \neq j} I_{j'}} \left\|X_{i(j)} u_i\right\|_2^2 - \frac{1}{\prod_{j' \neq j} I_{j'}} \mathbb{E}\left\{\left\|X_{i(j)} u_i\right\|_2^2\right\}\right| \geq s\right) \\ &\qquad \leq 2\exp\left\{-C_2 \prod_{j' \neq j} I_{j'} \cdot \min\left(\frac{s^2}{K^4}, \frac{s}{K^2}\right)\right\},\end{aligned}$$

where $K = \left\|x_{i(j)k}^T u\right\|_{\Psi_2}^2 = \kappa_0^2$. Define event $\Gamma(s) := \left\{\frac{1}{\prod_{j' \neq j} I_{j'}} \left\|X_{i(j)} u_i\right\|_2^2 \leq 2K + s\right\}$. Due to

$$\frac{1}{\prod_{j' \neq j} I_{j'}} \sum_{k=1}^{\prod_{j' \neq j} I_{j'}} \mathbb{E}\left\{\left\|X_{i(j)} u_i\right\|_2^2\right\} = \mathbb{E}\left\{\left(x_{i(j)1}^T u_i\right)^2\right\} \leq 2K,$$

under the condition of $\Gamma(s)$, it can be obtained from the total probability formula that

$$\begin{aligned}
&Pr\left(\left|\frac{1}{\prod_{j'\neq j} I_{j'}}\left\langle X_{i(j)}u_i, h'_\alpha\left(y_i-\langle\mathcal{B}^*_\alpha,\mathcal{X}_i\rangle\right)v_i\right\rangle\right|\geq t\right)\\
&\leq Pr\left(\left|\frac{1}{\prod_{j'\neq j} I_{j'}}\left\langle X_{i(j)}u_i, h'_\alpha\left(y_i-\langle\mathcal{B}^*_\alpha,\mathcal{X}_i\rangle\right)v_i\right\rangle\right|\geq t\middle|\Gamma(s)\right)+Pr(\Gamma^c(s))\\
&\leq 2\cdot\exp\left\{-\frac{C_1\left(\prod_{j'\neq j} I_{j'}\right)t^2}{\alpha^2(2K+s)}\right\}+2\cdot\exp\left\{-\frac{C_2\prod_{j'\neq j} I_{j'}}{K^2}\cdot\min\left(\frac{s^2}{K^2},s\right)\right\}\\
&\leq 2\cdot\exp\left\{-\frac{C_1\left(\prod_{j'\neq j} I_{j'}\right)t^2}{\alpha^2(K^2+1+s)}\right\}+2\cdot\exp\left\{-\frac{C_2\prod_{j'\neq j} I_{j'}}{K^2}\cdot\min\left(\frac{s^2}{K^2+1},s\right)\right\}.
\end{aligned}$$

Take $s=\sqrt{\frac{C_1}{2C_2}\frac{Kt}{\alpha}}$, yielding

$$\begin{aligned}
&Pr\left(\left|\frac{1}{\prod_{j'\neq j} I_{j'}}\left\langle X_{i(j)}u_i, h'_\alpha(y_i-\langle\mathcal{B}^*_\alpha,\mathcal{X}_i\rangle)v_i\right\rangle\right|\geq t\right)\\
&\leq\begin{cases}
4\cdot\exp\left\{-\dfrac{C_1\left(\prod_{j'\neq j} I_{j'}\right)t^2}{2\alpha^2(K^2+1)}\right\}, & t\leq\left(K+K^{-1}\right)\alpha\sqrt{\dfrac{2C_2}{C_1}},\\
4\cdot\exp\left\{-\sqrt{\dfrac{C_1C_2}{2}}\dfrac{\left(\prod_{j'\neq j} I_{j'}\right)t}{K\alpha}\right\}, & t>\left(K+K^{-1}\right)\alpha\sqrt{\dfrac{2C_2}{C_1}}.
\end{cases}
\end{aligned}$$

Furthermore, we can get

$$\begin{aligned}
&Pr\left(2\|\nabla H_\alpha(\mathcal{B}^*_\alpha)\|_*^*\geq t\right)\\
&\leq 9^{I_j+\prod_{j'\neq j} I_{j'}}\cdot\sum_{i=1}^N Pr\left(\left|\left\langle X_{i(j)}u_i, h'_\alpha(y_i-\langle\mathcal{B}^*_\alpha,\mathcal{X}_i\rangle)v_i\right\rangle\right|\geq\frac{t}{2}\right)\\
&\leq\begin{cases}
4N\cdot 9^{I_j+\prod_{j'\neq j} I_{j'}}\cdot\exp\left\{-\dfrac{C_1t^2}{8\alpha^2(K^2+1)\left(\prod_{j'\neq j} I_{j'}\right)}\right\}, & t\leq\left(K+K^{-1}\right)\alpha\sqrt{\dfrac{2C_2}{C_1}},\\
4N\cdot 9^{I_j+\prod_{j'\neq j} I_{j'}}\cdot\exp\left\{-\sqrt{\dfrac{C_1C_2}{8}}\dfrac{t}{K\alpha}\right\}, & t>\left(K+K^{-1}\right)\alpha\sqrt{\dfrac{2C_2}{C_1}}.
\end{cases}
\end{aligned}$$

Take $t=C'\alpha\kappa_0\sqrt{\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}}$, when $N\geq C''\left(I_j+\prod_{j'\neq j} I_{j'}\right)$, $Pr\left(2\|\nabla H_\alpha(\mathcal{B}^*_\alpha)\|_*^*\geq t\right)\to 0$, where C' and C'' are absolute constants. That means, there are absolute constants c_1, c_2 making the following formula true

$$Pr\left(2\|\nabla H_\alpha(\mathcal{B}^*_\alpha)\|_*^*\geq C'\alpha\kappa_0\sqrt{\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}}\right)\leq c_1\exp\left\{-c_2\left(\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}\right)\right\}.$$

This completes the proof. $\square$

From Lemma 3.8, we can get a reasonable value of λ, $C'\alpha\kappa_0\sqrt{\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}}$, which depends on the parameters α, κ_0 and $\sqrt{\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}}$. Combining this value and Theorems 3.4 – 3.7 lead to the following main result.

Theorem 3.9. *Under the conditions of Theorems* 3.4 *and* 3.7*, with* $\lambda = C'\alpha\kappa_0\sqrt{\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}}$*, we have the following bound*

$$\|\hat{\mathcal{B}}-\mathcal{B}^*\|_F \le C_1\sqrt{\kappa_u}\kappa_l^{-1}\left(\kappa_0^k+\sqrt{M_k}\right)\alpha^{1-k}+\left(C_2\kappa_2R_1N^{-1}log\left(\sum_{d=1}^{M}I_d\right)+C_3\lambda\right)\kappa_1^{-1}M^{-1}\sum_{d=1}^{M}\sqrt{r_d}$$

with high probability.

Remark 3.10. When $\alpha\to\infty$, problem (2.7) becomes the low-rank regularized least squares regression, which has no approximation error. Theorem 3.7 implies that the upper bound of $\|\hat{\mathcal{B}}-\mathcal{B}_\alpha^*\|_F$ is controlled by $\sum_{d=1}^{M}\sqrt{r_d}$, which is the same as that in [26, Theorem 1] with $s=0$ and $\alpha_k=\frac{1}{M}$. And Theorem 3.9 also indicates that the estimation can be robustified by choosing α if ϵ is heavy-tailed.

Remark 3.11. From the value of λ, it can be seen that λ is proportional to $\sqrt{\frac{I_j+\prod_{j'\neq j} I_{j'}}{N}}$. So this value will become larger as the dimension of tensor data increases. It makes the low-rank constraint effect of problem (2.7) stronger, so as to get a lower rank estimator to achieve the purpose of data dimension reduction.

$\boxed{4}$ ADMM-Based Algorithm

Introducing auxiliary tensors $\mathcal{Z}_d\in\mathbb{R}^{I_1\times\cdots\times I_M}$, $\forall d\in[M]$ and a vector $t\in\mathbb{R}^N$, problem (2.7) can be equivalently rewritten as

$$\begin{aligned}
&\min_{\mathcal{B},\{\mathcal{Z}_d\}_{d=1}^M,t} \frac{1}{N}\sum_{i=1}^{N}h_\alpha(t_i)+\lambda\cdot\frac{1}{M}\sum_{d=1}^{M}\|\mathcal{Z}_{d(d)}\|_*,\\
&s.t. \qquad \mathcal{B}=\mathcal{Z}_d,\quad \forall d\in[M],\\
&\qquad y_i-\langle\mathcal{B},\mathcal{X}_i\rangle=t_i,\quad \forall i\in[N].
\end{aligned}\tag{4.1}$$

The augmented Lagrangian function associated with the problem (4.1) can be written as

$$\begin{aligned}
&\mathcal{L}_\rho\left(\mathcal{B},\{\mathcal{Z}_d\}_{d=1}^M,t;\{\mathcal{Q}_d\}_{d=1}^M,r\right)\\
&=\frac{1}{N}\sum_{i=1}^{N}h_\alpha(t_i)+\frac{\lambda}{M}\sum_{d=1}^{M}\|\mathcal{Z}_{d(d)}\|_*+\sum_{d=1}^{M}\langle\mathcal{Q}_d,\mathcal{Z}_d-\mathcal{B}\rangle+\frac{\rho}{2}\sum_{d=1}^{M}\|\mathcal{Z}_d-\mathcal{B}\|_F^2\\
&\quad+\sum_{i=1}^{N}r_i\left(t_i-y_i+\langle\mathcal{B},\mathcal{X}_i\rangle\right)+\frac{\rho}{2}\sum_{i=1}^{N}\left(t_i-y_i+\langle\mathcal{B},\mathcal{X}_i\rangle\right)^2,
\end{aligned}$$

where $\rho > 0$ is the augmented Lagrangian parameter, and $\{Q_d\}_{d=1}^M, r$ are Lagrangian multipliers. The iterative scheme of ADMM is described as:

$$\begin{cases} \mathcal{B}^{k+1} &= \arg\min_{\mathcal{B}} \left\{ \mathcal{L}_\rho(\mathcal{B}, \{\mathcal{Z}_d^k\}_{d=1}^M, t^k; \{\mathcal{Q}_d^k\}_{d=1}^M, r^k) \right\}, \\ \left(\left\{ \mathcal{Z}_d^{k+1} \right\}_{d=1}^M, t^{k+1} \right) &= \arg\min_{\{\mathcal{Z}_d\}_{d=1}^M, t} \left\{ \mathcal{L}_\rho(\mathcal{B}^{k+1}, \{\mathcal{Z}_d\}_{d=1}^M, t; \{\mathcal{Q}_d^k\}_{d=1}^M, r^k) \right\}, \\ \mathcal{Q}_d^{k+1} &= \mathcal{Q}_d^k + \tau\rho \left(\mathcal{Z}_d^{k+1} - \mathcal{B}^{k+1} \right), \quad d \in [M], \\ r_i^{k+1} &= r_i^k + \tau\rho \left(t_i^{k+1} - y_i + \langle \mathcal{B}^{k+1}, \mathcal{X}_i \rangle \right), \quad i \in [N]. \end{cases} \tag{4.2}$$

where $\tau > 0$ is referred as the dual step size, with a typical choice $\tau = 1.618$ which is adopted in this paper. For the first subproblem in (4.2), by vectorizing all the tensors, e.g., $x_i = \mathtt{vec}(\mathcal{X}_i)$ and denote $X = (x_1, \ \cdots, x_M)^T$, we can get the following closed form solution

$$\mathcal{B}^{k+1} = \mathtt{vtt} \left(\left(X^T X + M\mathbb{I} \right)^{-1} \left(\sum_{d=1}^M \left(z_d^k + \frac{q_d^k}{\rho} \right) - X^T \left(t^k - y + \frac{r^k}{\rho} \right) \right) \right), \tag{4.3}$$

where $z_d^k = \mathtt{vec}(\mathcal{Z}_d^k)$, $q_d^k = \mathtt{vec}(\mathcal{Q}_d^k)$ and $\mathtt{vtt}$ is the inverse operator of $\mathtt{vec}$ in the underlying spaces. For any given $d \in [M]$, we can get the $\mathcal{Z}_d$-update by employing the singular value thresholding in [6]

$$\mathcal{Z}_d^{k+1} = fold_d \left[\mathtt{Prox}_{\frac{\lambda}{M\rho}\|\cdot\|_*} \left(B_{(d)}^{k+1} - \frac{Q_{d(d)}^k}{\rho} \right) \right]. \tag{4.4}$$

For any given $i \in [N]$, we can employ the proximal operator as described in Lemma 2.1 to get the t_i-update by

$$t_i^{k+1} = \mathtt{Prox}_{\frac{1}{N\rho} h_\alpha(\cdot)} \left[y_i - \langle \mathcal{B}^{k+1}, \mathcal{X}_i \rangle - \frac{r_i^k}{\rho} \right]. \tag{4.5}$$

The framework of ADMM for solving problem (4.1) is then summarized in Algorithm 1.

Algorithm 1 ADMM for Solving Problem (4.1)

Input: The observations $\{(\mathcal{X}_i, y_i) : i \in [N]\}$ and parameters ρ, λ, τ, α.
Output: $\mathcal{B}^k$.
Step 1. Initialize $\left(\mathcal{B}^0, \{\mathcal{Z}_d^0\}_{d=1}^M, \{t_i^0\}_{i=1}^N, \{\mathcal{Q}_d^0\}_{d=1}^M, \{r_i^0\}_{i=1}^N\right)$ to be zero, and $k = 0$;
Step 2. Compute $\left(\mathcal{B}^{k+1}, \{\mathcal{Z}_d^{k+1}\}_{d=1}^M, \{t_i^{k+1}\}_{i=1}^N, \{\mathcal{Q}_d^{k+1}\}_{d=1}^M, \{r_i^{k+1}\}_{i=1}^N\right)$ by (4.3), (4.4), (4.5) and (4.2), respectively;
Step 3. Set $k = k + 1$. If some stopping criterion is met, then stop; Otherwise, go to Step 2.

Stopping Criterion. Applying the classical convex optimization theory, we adopt the relative primal infeasibility and relative dual infeasibility, defined as below, to measure the quality of the approximate solution:

$$\begin{aligned} \eta_P &:= \max\left\{ \eta_{Z_1}, \ldots, \eta_{Z_M}, \eta_{r_1}, \ldots, \eta_{r_N}, \eta_{Q_1}, \ldots, \eta_{Q_M}, \eta_{t_1}, \ldots, \eta_{t_N} \right\}, \\ \eta_D &:= \frac{\left\| - \sum_{d=1}^M \mathcal{Q}_d^k + \sum_{i=1}^N r_i^k \mathcal{X}_i \right\|_F}{1 + \|\mathcal{B}^k\|_F}, \end{aligned}$$

where

$$\eta_{Z_d} := \frac{\left\| \mathcal{Z}_d^k - fold_d \left[\mathtt{Prox}_{\frac{\lambda}{M}\|\cdot\|_*} \left(Z_{d(d)}^k - Q_{d(d)}^k \right) \right] \right\|_F}{1 + \left\| \mathcal{Z}_d^k \right\|_F}, \ \eta_{Q_d} := \frac{\left\| \mathcal{Z}_d^k - \mathcal{B}^k \right\|_F}{1 + \left\| \mathcal{Q}_d^k \right\|_F}, \ \text{for } d \in [M],$$

and

$$\eta_{r_i} := \frac{\left| \frac{1}{N} \nabla h_\alpha \left(t_i^k \right) + r_i^k \right|}{1 + \left| t_i^k \right|}, \ \eta_{t_i} := \frac{\left| t_i^k - y_i + \langle \mathcal{B}^k, \mathcal{X}_i \rangle \right|}{1 + \left| r_i^k \right|}, \ \text{for } i \in [N].$$

It is reasonable to terminate Algorithm 1 if $\max\{\eta_p, \eta_D\} \leq \varepsilon$, where $\varepsilon \geq 0$ is a prescribed accuracy parameter.

Global Convergence. The global convergence of Algorithm 1 follows readily from the classical two-block ADMM for convex program, since we can treat $\mathcal{B}$ and $\left(\{\mathcal{Z}_d\}_{d=1}^M, t\right)$ as these two blocks. The proximal ADMM scheme (see, e.g., [12]) can also be employed to approximately update $\mathcal{B}^k$ to reduce the computational cost from (4.3). The global convergence in this regime is also guaranteed by [12, Appendix B.2].

Computational Complexity. Let $n = \prod_{d=1}^M I_d$. The main computation in each iteration of Algorithm 1 comes from the updates for $\mathcal{B}$, $\{\mathcal{Z}_d\}_{d=1}^M$ and $\left\{t, \{Q_d\}_{d=1}^M, r\right\}$, which are of order $O\left(n^3\right)$, $O\left(Mn \min\left(I_d, \prod_{d' \neq d} I'_d\right)\right)$ and $O\left((M+N)\,n\right)$, respectively. Hence, the per-iteration computational complexity of Algorithm 1 is of order $O(n^3)$ in high-dimensional regression settings, dominated by the matrix inverse in (4.3). Fortunately, as one can see, the involved matrix inversion remains the same in the entire iteration process, which can be computed before the main loop. Additionally, the conjugate gradient (CG) method can be called to handle the underlying linear system for an approximate update for $\mathcal{B}$.

5 Numerical Experiments

In this section, we conduct numerical experiments to examine the effectiveness of the NNTH estimator and to evaluate the performance of our proposed ADMM algorithm. All numerical experiments are implemented in MATLAB (R2018b), running on a laptop with Intel Core i5 CPU (1.867GHz) and 8 GB RAM.

5.1 Simulation Studies

We randomly generate the ground-truth coefficient tensor $\mathcal{B}^* = b \cdot \mathcal{C} \times_1 M_1 \times_2 M_2 \times_3 M_3$, where $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, $M_1 \in \mathbb{R}^{I_1 \times r_1}$, $M_2 \in \mathbb{R}^{I_2 \times r_2}$, $M_3 \in \mathbb{R}^{I_3 \times r_3}$ are with element-wise *i.i.d.* standard Gaussian distribution, and $b > 0$ is the signal strength. The responses are generated by $y_i = \langle \mathcal{B}, \mathcal{X}_i \rangle + \varepsilon_i$, $i \in [N]$, where $\mathcal{X}_i \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ has standard normal entries, and random errors ε_i's are generated from the following three distributions: (i) normal errors with mean 0 and variance 4 ($N(0,4)$); (ii) log-normal distribution(LogNormal), $\varepsilon = \exp(1 + 5Z)$, where Z is the standard normal distribution; (iii) Weibull distribution with shape parameter 0.2 and scale parameter 0.7.

Set $I_1 = 20$, $I_2 = 10$, $I_3 = 30$ with a variety of low Tucker ranks ($(r_1, r_2, r_3) = (1,1,2), (2,1,3), (2,2,2)$) and signal strengths ($b = 1, 2, 5$). For each scenario, our simulated data consist of a training set of 1000 samples and an independent testing set of 100 samples. Hyperparameters (λ, α) in NNTH estimation will be determined via 5-fold cross validation on the training set over a grid of (λ, α)'s with varying $\lambda \in \{10^{-2}, 5\times10^{-2}, 10^{-1}, \cdots, 5\times10^{2}, 10^{3}\}$

and $\alpha \in \{0.5, 1.345, 2, 3.45, 5\}$ (The choice of $\alpha = 1.345$ was observed to gain promising performance in [19]). Parameters with the minimum mean square error (MSE) on the verification set will be chosen. Here MSE of a given estimator $\hat{\mathcal{B}}$ is defined by MSE $= \frac{\|\hat{\mathcal{B}}-\mathcal{B}^*\|_F^2}{I_1 \times I_2 \times I_3}$. The parameter λ in all comparing approaches in the sequel will be chosen in the same fashion.

Comparisons to other approaches including the nuclear norm regularized tensor least squares regression (NNTLS) [37], Lasso [40] and Elastic Net(ENet) [46] are carried out. Each simulation is based on 50 independent replications, and the average results are depicted in Tables 2, 3 and 4, with the best results highlighted in bold and the second-best ones underlined.

With normal errors which are symmetric and light-tailed, Table 2 indicates that, the NNTLS estimator reasonably gains the best performance due to the low-rank promoting term by tensor nuclear norm regularization, and the least squares loss tailored for normal errors. NNTH reaches very competitive performances to NNTLS, and outperform Lasso and ENet with efforts on entrywise sparsity.

With asymmetric and heavy-tailed errors, e.g., the Weibull and LogNormal errors, Tables 3 and 4 illustrate the significant superiority of NNTH in terms of MSE for all the testing instances. In particular, Table 4 shows that NNTH estimator has the greatest advantage in dealing with LogNormal distribution error models, e.g., the estimation errors of NNTH are nearly 1/10000 or even 1/100000 of those generated by NNTLS. In both cases, NNTLS, Lasso, ENet and Ridge estimators do not perform well owing to the sensitivity of secondary loss to outliers. This illustrates the merit of our proposed tensor Huber regression model. It is noteworthy that the accuracy of NNTH estimators for all testing cases are around decreases with the increase of the signal strength or the Tucker rank complexity $\sum_{d=1}^{M} \sqrt{r_d}$. Such a phenomenon can be explained by Theorem 3.9.

As for all testing instances shown in Tables 2, 3 and 4, NNTH estimator achieves the MSE mostly of order 10^{-3}, which reflects a promisingly robust behavior comparing to other approaches in the simulation studies.

Table 2: The performance of methods for normal error model.

	r=(1,1,2)				r=(2,1,3)				r=(2,2,2)			
	NNTH	NNTLS	Lasso	ENet	NNTH	NNTLS	Lasso	ENet	NNTH	NNTLS	Lasso	ENet
b=1	1.49e-3	**1.48e-3**	1.98e-3	1.69e-3	2.14e-3	**2.12e-3**	2.23e-3	2.18e-3	2.18e-3	**2.17e-3**	2.28e-3	2.27e-3
b=2	2.87e-3	**2.85e-3**	3.01e-3	2.97e-3	3.84e-3	**3.82e-3**	4.16e-3	3.94e-3	4.55e-3	**4.51e-3**	4.79e-3	4.67e-3
b=5	8.16e-3	**8.10e-3**	8.66e-3	8.55e-3	7.78e-3	**7.73e-3**	8.11e-3	7.94e-3	1.15e-2	**1.14e-2**	1.19e-2	1.18e-2

Table 3: The performance of methods for Weibull error model.

	r=(1,1,2)				r=(2,1,3)				r=(2,2,2)			
	NNTH	NNTLS	Lasso	ENet	NNTH	NNTLS	Lasso	ENet	NNTH	NNTLS	Lasso	ENet
b=1	**1.28e-3**	1.94e-2	6.75e-1	4.71e-1	**1.98e-3**	2.21e-2	6.94e-2	5.17e-2	**2.11e-3**	2.66e-2	5.92e-2	4.65e-2
b=2	**1.91e-3**	2.42e-2	2.45e-2	1.78e-1	**2.70e-3**	3.42e-2	1.05e-1	8.06e-2	**3.53e-3**	3.61e-2	4.86e-3	4.50e-3
b=5	**5.05e-3**	3.03e-2	1.12e-1	7.75e-2	**7.27e-3**	4.32e-2	2.65e-2	1.97e-2	**8.57e-3**	4.86e-2	2.40e-1	1.76e-1

Table 4: The performance of methods for LogNormal error model.

	r=(1,1,2)				r=(2,1,3)				r=(2,2,2)			
	NNTH	NNTLS	Lasso	ENet	NNTH	NNTLS	Lasso	ENet	NNTH	NNTLS	Lasso	ENet
b=1	**1.74e-3**	3.00e+1	1.96e+2	1.34e+2	**1.97e-3**	7.13e+1	1.37e+1	9.30e+0	**2.47e-3**	2.76e+2	1.11e-1	7.93e-2
b=2	**2.53e-3**	1.14e+2	1.54e+2	1.10e+2	**3.20e-3**	1.84e+2	8.81e+0	6.35e+0	**4.20e-3**	5.16e+2	9.56e+0	6.82e+0
b=5	**8.47e-3**	4.54e+2	1.40e+0	9.29e-1	**8.82e-3**	1.47e+3	6.82e+1	4.72e+1	**9.98e-3**	4.16e+3	8.59e-1	6.20e-1

5.2 Analysis of CIFAR-10 Dataset

In this subsection, we apply our model to classify the CIFAR-10 dataset [24], whose 3D data size is $32 \times 32 \times 3$ (of total $3,072$ voxels). We randomly select five class pairs to do binary classification. Without overlap, 100 samples are randomly selected from each class set for model training and testing.

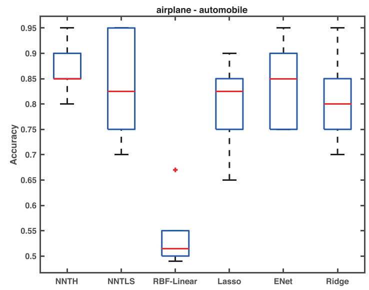
We randomly divide the two class datasets into training set $(\mathcal{X}_{training}, y_{training})$ with 180 samples and test set $(\mathcal{X}_{test}, y_{test})$ with 20 samples. Firstly, the training set $(\mathcal{X}_{training}, y_{training})$ is used to fit the model, and the estimator $\hat{\mathcal{B}}_{training}$ is obtained. Then we use it to predict and classify on the test set. The classification accuracy (ACC) will be adopted to measure the performance of our method and other comparing methods including NNTLS [37], RBF-Linear algorithm [8], Lasso [40], ENet [46], and Ridge [16]. In order to reduce the impact of data set segmentation as much as possible, we randomly segment the data for 10 times and use 10-fold cross validation. We use the mean and variance of these 10 numerical results to reflect the effectiveness and robustness of all the methods. Performances are summarized in Table 5 and Figure 2.

Table 5 shows the classification accuracy of all methods, with the best results highlighted in bold and the second-best ones underlined. Among these six approaches, NNTH gives the best classification accuracy in most cases, and the average accuracy of NNTH estimator (77.70%) is 3.10% higher than the second-best NNTLS estimator (74.60%), both of which outperform vector-based approaches including Lasso (69.60%), ENet (70.80%) and Ridge (73.20%). Figure 2 shows boxplots of ACC for all methods, which indicates the robustness and high accuracy of NNTH in handling 3D data classification. Some selected instances in the test sets by NNTH is presented in Figure 3 where misclassified images are marked in red boxes.

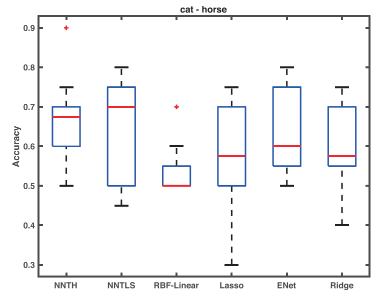
Table 5: Numerical results on the CIFAR-10 dataset.

Class pair		NNTH	NNTLS	RBF-Linear	Lasso	ENet	Ridge
‘airplane’, ‘automobile’	Avg(ACC)	**86.50**	83.00	54.60	80.50	<u>83.50</u>	81.00
	Std(ACC)	0.0530	0.1085	0.0685	0.0832	0.0709	0.0810
‘cat’, ‘horse’	Avg(ACC)	**66.50**	<u>65.00</u>	54.00	58.00	63.50	59.50
	Std(ACC)	0.1132	0.1269	0.0658	0.1418	0.1081	0.1066
‘airplane’, ‘truck’	Avg(ACC)	**85.50**	82.00	68.50	78.00	81.00	<u>85.00</u>
	Std(ACC)	0.0497	0.0856	0.1226	0.0632	0.0568	0.0577
‘automobile’, ‘cat’	Avg(ACC)	<u>72.50</u>	71.50	**73.50**	71.50	69.00	72.00
	Std(ACC)	0.0755	0.1029	0.0973	0.0474	0.1174	0.0753
‘cat’, ‘deer’	Avg(ACC)	**77.50**	<u>71.50</u>	64.00	60.00	57.00	68.50
	Std(ACC)	0.0920	0.1055	0.0775	0.0972	0.1059	0.1510
	$\overline{\text{Avg(ACC)}}$	**77.70**	<u>74.60</u>	62.92	69.60	70.80	73.20

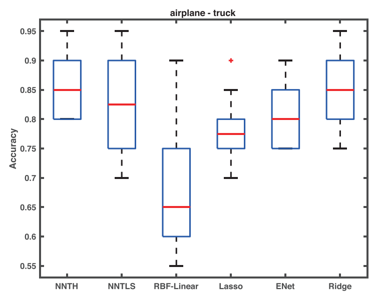
To summarize, the real data analysis confirms the applicability of Huber loss function, and also shows that the tensor nuclear norm regularizer has a good ability of low-rank structure modeling in 3D real data.



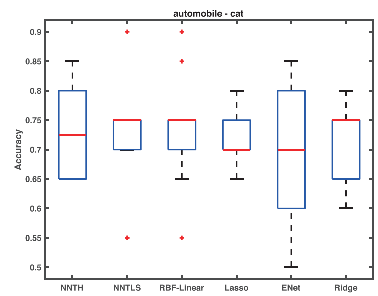
(a) 'airplane','automobile'



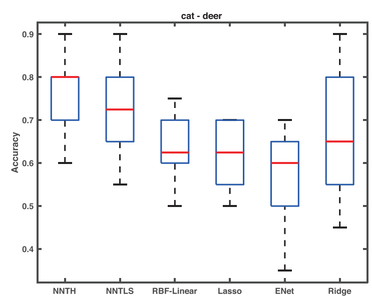
(b) 'cat', 'horse'



(c) 'airplane', 'truck'



(d) 'automobile', 'cat'



(e) 'cat', 'deer'

Figure 2: The boxplots of ACC for the CIFAR-10 dataset.

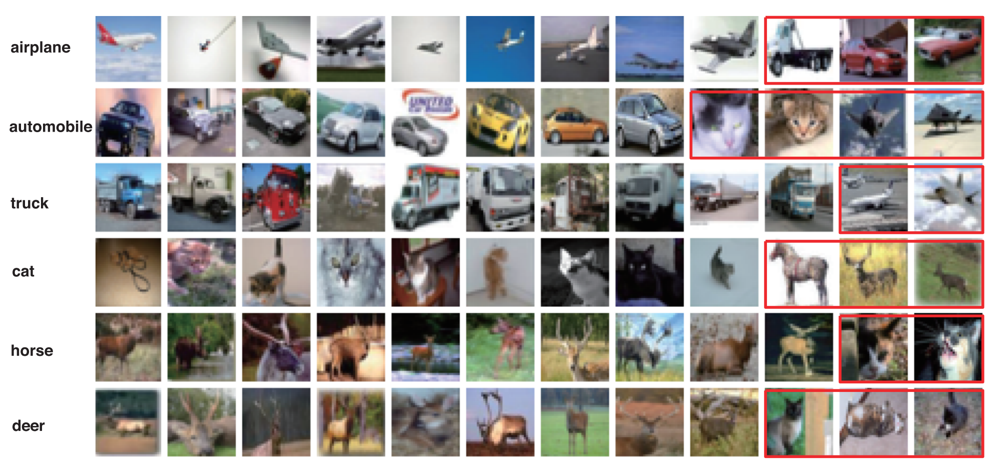


Figure 3: Examples of image classification results by NNTH.

$\boxed{6}$ Conclusions

In this paper, we have considered the nuclear norm regularized tensor Huber regression (NNTH) method, which can effectively handle the tensor data with low-rank structure and outliers/heavy-tailed errors. By virtue of decomposability of nuclear norm and restricted strong convexity of Huber loss function, the upper bound of estimation error has been established in the sense of Frobenius norm. An ADMM algorithm has been designed and the numerical results have verified the effectiveness of the proposed NNTH method. Besides the nuclear norm regularization for the tensor low-rankness, it would be of significance for future research to develop the robust low-rank Huber tensor regression methods based on tensor decomposition, such as Tucker decomposition [45, 27] and CANDECOMP/PARAFAC (CP) decomposition [15, 39], for further dimension reduction.

References

[1] D. Akkaya and M. Ç. Pınar, Minimizers of sparsity regularized Huber loss function, *J. Optim. Theory Appl.* 187 (2020) 205–233.

[2] V. Barnett and T. Lewis, *Outliers in Statistical Data* (3rd ed) Chichester: John Wiley & Sons Inc, 1994.

[3] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Space*, Springer, New York, 2011.

[4] X. Bi, X. Tang, Y. Yuan, Y. Zhang and A. Qu, Tensors in statistics, *Annu. Rev. Stat. Appl.* 8 (2021) 345–368.

[5] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*, Springer, New York, 2011.

[6] J. Cai, E. J. Candès and Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optim.* 20 (2010) 1956–1982.

[7] B. Chen, W. Zhai and Z. Huang, Low-rank elastic-net regularized multivariate Huber regression model, *Appl. Math. Model.* 87 (2020) 571–583.

[8] C. Chen, K. Batselier, W. Yu and N. Wong, Kernelized support tensor train machines, *arXiv preprint arXiv:2001.00360v1*, 2020.

[9] A. Elsener and S. van de Geer, Robust low-rank matrix estimation, *Ann. Statist.* 46 (2018) 3481–3509.

[10] J. Fan, Q. Li and Y. Wang, Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions, *J. R. Stat. Soc. Ser. B Stata. Methodol.* 79 (2017) 247–265.

[11] X. Fang, K. Paynabar and N. Gebraeel, Image-based prognostics using penalized tensor regression, *IEEE Trans. Pattern Anal. Mach. Intell.* 61 (2017) 369–384.

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

[13] S. Foss, D. Korshunov and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions*, New York: Springer, 2011.

[14] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, 1986.

[15] B. Hao, A. Zhang and G. Cheng, Sparse and low-rank tensor estimation via cubic sketchings, *IEEE Trans. Inf. Theory* 66 (2020) 5927–5964.

[16] A.E. Hoerl and R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12 (2012) 55–67.

[17] M. Hou and B. Chaib-draa, Hierarchical Tucker tensor regression: Application to brain imaging data analysis, in: *IEEE International Conference on Image Processing (ICIP)*, 2015.

[18] P.J. Huber, Robust estimation of a location parameter, *Ann. Math. Statist.* 35 (1964) 73–101.

[19] P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[20] C.M. Hurvich and C.-L. Tsai, Model selection for least absolute deviations regression in small samples, *Statist. Probabil. Lett.* 9 (2008) 259–265.

[21] R.W. Koenker and G. Bassett, Regression quantile, *Econometrica* 46 (1978) 33–50.

[22] T.G. Kolda and B.W. Bader, Tensor decompositions and applications, *SIAM Rev.* 51 (2009) 455–500.

[23] D. Kong, B. An, J. Zhang and H. Zhu, L2RM: Low-rank linear regression models for high-dimensional matrix responses, *J. Amer. Statist. Assoc.* 115 (2020) 403–424.

[24] A. Krizhevsky and G. Hinton, Learning multiple layers of features from tiny images, *Citeseer, Technical Report*, 2009.

[25] S. Lambert-Lacroix and L. Zwald, Robust regression through the Huber's criterion and adaptive lasso penalty, *Electron. J. Statist.* 5 (2011) 1015–1053.

[26] X. Li, A. Wang, J. Lu and Z. Tang, Statistical performance of convex low-rank and sparse tensor recovery, in: *IAPR Asian Conference on Pattern Recognition*, volume 1, 2017 pp. 524–529.

[27] X. Li, D. Xu, H. Zhou and L. Li, Tucker tensor regression and neuroimaging analysis, *Statist. Biosci.* 10 (2018) 520–545.

[28] H. Lian, Learning rate for convex support tensor machines, *IEEE Trans. Neur. Net. Lear. Systems* 32 (2021) 3755–3760.

[29] J. Liu, P. Musialski, P. Wonka and J. Ye, Tensor Completion for Estimating Missing Values in Visual Data, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (2013) 208–220.

[30] Y. Liu, P. Zeng and L. Lin, Degrees of freedom for regularized regression with Huber loss and linear constraints, *Statist. Pap.*, 2020. https://doi.org/10.1007/s00362-020-01192-2.

[31] H. Lu, K.N. Plataniotis and A.N. Venetsanopoulos. MPCA: Multilinear principal component analysis of tensor objects, *IEEE Trans. Neur. Net.* 19 (2008) 18–39.

[32] S.N. Negahban, P. Ravikumar, M.J. Wainwright and B. Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statist. Sci.* 27 (2010) 538–557.

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

[34] A. Novikov, D. Podoprikhin, A. Osokin and D. Vetrov, Tensorizing neural networks, in: *International Conference on Neural Information Processing Systems (NIPS)*, 2015 pp. 442–450.

[35] L. Qi and Z. Luo, *Tensor Analysis: Spectral Theory and Special Tensors*, t SIAM Press, 2017.

[36] G. Raskutti, M. Yuan and H. Chen, Convex regularization for high-dimensional multiresponse tensor regression, *Ann. Statist.* 47 (2019) 1554–1584.

[37] X. Song and H. Lu, Multilinear regression for embedded feature selection with application to fMRI analysis, in: *AAAI Conference on Artificial Intelligence*, volume 17, 2017 pp. 2562–2568.

[38] Q. Sun, Wen. Zhou and J. Fan, Adaptive Huber regression, *J. Amer. Statist. Assoc.* 115 (2020) 254–265.

[39] W.W. Sun and L. Li, STORE: Sparse tensor response regression and neuroimaging analysis, *J. Mach. Lear. Res.* 18 (2017) 1–37.

[40] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Stata. Methodol.* 73(1996) 273–282.

[41] R. Tomioka, K. Hayashi and H. Kashima, Estimation of low-rank tensors via convex optimization, *arXiv preprint arXiv:1010.0789*, 2010.

[42] R. Tomioka, T. Suzuki, K. Hayashi and H. Kashima, Statistical performance of convex tensor decomposition, in: *Advances in Neural Information Processing Systems (NIPS)*, 2011 pp. 972–980.

[43] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, *Cambridge University Press*, volume 47, 2018.

[44] S. Yan, D. Xu, Q. Yang and L. Zhang, Multilinear discriminant analysis for face recognition, *IEEE Trans. Image Process.* 16 (2007) 212–220.

[45] A. Zhang, Y. Luo, G. Raskutti and M. Yuan, ISLET: Fast and optimal low-rank tensor regression via importance sketching, *SIAM J. Math. Data Sci.* 2 (2020) 444–479.

[46] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stata. Methodol.* 67 (2005) 301–320.

Manuscript received 22 May 2021
revised 11 August 2021
accepted for publication 27 September 2021

Yangxin Wei
Department of Mathematics
Beijing Jiaotong University
Beijing, 100044, China
E-mail address: 19121637@bjtu.edu.cn

Ziyan Luo
Department of Mathematics
Beijing Jiaotong University
Beijing, 100044, China
E-mail address: zyluo@bjtu.edu.cn

Yang Chen
Department of Mathematics
Beijing Jiaotong University
Beijing, 100044, China
E-mail address: 17121619@bjtu.edu.cn