Half-quadratic alternating direction method of multipliers for robust orthogonal tensor approximation

Yuning Yang¹ · Yunlong Feng²

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Higher-order tensor canonical polyadic decomposition (CPD) with one or more of the latent factor matrices being columnwisely orthonormal has been well studied in recent years. However, most existing models penalize the noises, if occurring, by employing the least squares loss, which may be sensitive to non-Gaussian noise or outliers, leading to bias estimates of the latent factors. In this paper, we derive a robust orthogonal tensor CPD model with Cauchy loss, which is resistant to heavy-tailed noise such as the Cauchy noise, or outliers. By exploring the half-quadratic property of the model, we develop the so-called half-quadratic alternating direction method of multipliers (HQ-ADMM) to solve the model. Each subproblem involved in HQ-ADMM admits a closed-form solution. Thanks to some nice properties of the Cauchy loss, we show that the whole sequence generated by the algorithm globally converges to a stationary point of the problem under consideration. Numerical experiments on synthetic and real data demonstrate the effectiveness of the proposed model and algorithm.

Keywords  Tensor · Canonical polyadic decomposition · Robust · Cauchy · HQ-ADMM

Communicated by: Guoyin Li

This article belongs to the Topical Collection: Mathematics of Computation and Optimisation
Guest Editors: Jerome Droniou, Andrew Eberhard, Guoyin Li, Russell Luke, Thanh Tran

✉ Yuning Yang
yyang@gxu.edu.cn

Yunlong Feng
ylfeng@albany.edu

¹ College of Mathematics and Information Science, Guangxi University, Nanning, 530004, China

² Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY, 12222, USA
1 Introduction

A tensor is a multidimensional array. Owing to its ability to represent data with intrinsically many dimensions, tensors draw much attention from the communities of signal processing, image processing, machine learning, etc. (see the surveys [7, 28, 45]). To understand the relationship behind the data tensor, decomposition tools are needed. In general, tensor decomposition aims at factorizing the data tensor into a set of lower-dimensional latent factors, where the factors can be vectors, matrices, or even tensors. Tensor canonical polyadic decomposition (CPD), which factorizes a tensor into a sum of component rank-1 tensors, is one of the most important tensor decomposition models. Tensor CPD finds applications in blind multiuser CDMA, blind source separation, and so on [45]. Different from matrix decompositions, tensor CPD is unique under quite mild conditions [28].

In some applications, one or more latent factors of the CPD are required to have orthonormal columns. For example, in linear image coding [44], one is given a set of data matrices of the same size; to explore their commonalities, one projects the matrices onto a latent lower-dimensional subspace, which can be represented by the Khatri-Rao product [28] of two columnwisely orthonormal matrices. This problem has been formulated as a third-order tensor CPD with two factor matrices having orthonormal columns. On the other hand, simultaneous foreground-background extraction and compression can also be formulated as a model of the same kind; this will be illustrated in Section 5. Other applications of CPD with orthonormal factors can be found in [8–10, 46, 49].

In reality, due to the NP-hardness of determining the tensor rank [20], and due to the presence of noise, tensor CPD model is rarely exact, and it is necessary to resort to an approximation scheme. To numerically solve the problem, one usually formulates it as an optimization problem with least squares loss–based objective and orthonormal constraints, and then applies a polar decomposition–based method to solve it [5, 18, 25, 35, 40, 48, 51, 53]. Other types of methods can be found in [11, 34, 42], just to name a few.

Although the optimization model mentioned above is effective in some circumstances, note that the least squares loss is not robust to non-Gaussian noise [26]. As a result, when the data tensor is contaminated by heavy-tailed noise, such as the Cauchy noise, or outliers, such least squares–based models may lead to bias estimates of the true latent factors, as having been observed in practice. This drawback of the least squares–based models motivates us to develop a new model that is robust to heavy-tailed noise or outliers.

In this work, from the maximum a posterior estimation, we derive a robust tensor CPD model where one or more latent factors have orthonormal columns. Such a model is based on the Cauchy loss, whose robustness comes from the redescending property [26] of the loss function. We then explore the half-quadratic property of the model, based on which, the half-quadratic alternating direction method of multipliers (HQ-ADMM) is proposed to solve the model. An advantage of HQ-ADMM
is that every subproblem involved in the algorithm admits a closed-form solution. Under a very mild assumption on the parameter, HQ-ADMM is proved to globally converge to a stationary point of the problem under consideration, owing to some nice properties of the Cauchy loss. In fact, the spirit of HQ-ADMM can be applied to solve other Cauchy loss–based machine learning and scientific computing problems (besides tensor problems), which will be remarked later in Section 3. Finally, we show via numerical experiments that the proposed model is resistant to heavy-tailed noise such as Cauchy noise, outliers, and also performs well with Gaussian noise.

The rest of the paper is organized as follows. The robust tensor approximation model is formulated in Section 2, with some quantitative properties given. The HQ-ADMM is developed in Section 3; the convergence analysis of HQ-ADMM is provided in Section 4. Numerical results are illustrated in Section 5. We end this paper in Section 6 with conclusions.

2 Problem formulation and the optimization model

Notations Vectors are written as boldface lowercase letters (x, y, . . .), matrices are denoted as italic capitals (A, B, . . .), and tensors are written as calligraphic capitals (A, B, . . .). \( \mathbb{R} \) denotes the real field. \( \mathbb{R}^{m \times n} \) denotes real matrices of dimension \( m \times n \) and \( \mathbb{R}^{n_1 \times \cdots \times n_d} \) denotes tensor space of size \( n_1 \times \cdots \times n_d \). The Frobenius norm, \( \| \cdot \|_F \), of a matrix or a tensor, is defined to be the square root of the sum of squares of all the entries. The inner product \( \langle \cdot, \cdot \rangle \) between a pair of matrices or tensors of the same size is given by the sum of entrywise product. \( \otimes \) denotes the outer product of two vectors.

Let \( A = (A_{i_1 \cdots i_d}) \in \mathbb{R}^{n_1 \times \cdots \times n_d} \) be a \( d \)th order observed data tensor. We consider the inexact CPD of \( A \), i.e., approximating \( A \) by a sum of rank-1 tensors:

\[
A = \sum_{i=1}^{R} \sigma_i \bigotimes_{j=1}^{d} u_{j,i} + N \in \mathbb{R}^{n_1 \times \cdots \times n_d},
\]

(2.1)

here \( u_{j,i} \in \mathbb{R}^{n_j} \), \( 1 \leq j \leq d \), \( \bigotimes_{j=1}^{d} u_{j,i} \) denotes the rank-1 tensor given by the outer product of \( u_{j,i} \)'s, \( \sigma_i \)'s are real scalars, \( R > 0 \) is a given integer, where usually \( R \) is such that \( R \leq \min\{n_1, \ldots, n_d\} \) for a possibly low-rank approximation, while \( N \) denotes the noisy tensor.

Denote \( U_j := \begin{bmatrix} u_{j,1}, \ldots, u_{j,R} \end{bmatrix} \in \mathbb{R}^{n_j \times R} \) and \( \sigma := [\sigma_1, \ldots, \sigma_R] \in \mathbb{R}^R \). Then, \( U_j \)'s are called the latent factor matrices of \( A \). Throughout this work, we follow [28] to write the sum of rank-1 terms as

\[
\sigma; U_1, \ldots, U_d := \sum_{i=1}^{R} \sigma_i \bigotimes_{j=1}^{d} u_{j,i};
\]

moreover, we write

\[
U := \{U_1, \ldots, U_d\} \text{ and } \sigma; U := \sigma; U_1, \ldots, U_d
\]

for short. In the sequel, we base our work on the following setup:
– One or more $U_j$’s are columnwisely orthonormal. Without loss of generality, we assume that the last $t$ $(1 \leq t \leq d)$ matrices are columnwisely orthonormal, i.e.,

$$U_j^TU_j = I, \ d - t + 1 \leq j \leq d,$$

where $I$ is an identity matrix of the proper size;
– The columns of the first $d - t$ matrices are normalized, i.e.,

$$\|u_{j,i}\| = 1, \ 1 \leq j \leq d - t, \ 1 \leq i \leq R;$$
– Entries of the noisy tensor $\mathcal{N}$ are i.i.d..

We immediately have the following proposition.

**Proposition 2.1** There holds

$$\left(\bigotimes_{j=1}^{d} u_{j,i_1}, \bigotimes_{j=1}^{d} u_{j,i_2}\right) = 0, \ i_1 \neq i_2, \text{ and}$$

$$\left\|\bigotimes_{j=1}^{d} u_{j,i}\right\|_F = 1, \ 1 \leq i \leq R.$$

Note that the constraints on $u_{j,i}$ and $U_j$ are all Stiefel manifolds $\text{st}(m,n) := \{P \in \mathbb{R}^{m \times n} | P^TP = I\}$. Therefore, in the following, we write the constraints on $u_{j,i}$ and $U_j$ as

$$u_{j,i} \in \text{st}(n_j, 1), \ 1 \leq j \leq d - t, \ 1 \leq i \leq R,$$

$$U_j \in \text{st}(n_j, R), \ d - t + 1 \leq j \leq R.$$

In the presence of the noisy term $\mathcal{N}$, it is natural to deal with (2.1) via solving the following optimization problem [5, 18, 48, 51, 53]:

$$\min_{\sigma, u_{j,i} \in \text{st}(n_j, 1), U_j \in \text{st}(n_j, R)} \left\|A - \left[\sigma; U\right]\right\|_F^2 = \sum_{i_1=1, \ldots, i_d=1}^{n_1, \ldots, n_d} \left(A_{i_1 \ldots i_d} - \left[\sigma; U\right]_{i_1 \ldots i_d}\right)^2.$$  \hspace{1cm} (2.2)

From a statistical estimation viewpoint, the above model is built upon the least squares loss $\ell_2(t) := t^2/2$, i.e., it employs the $\ell_2(\cdot)$ loss to deal with noise. However, it is commonly known that the estimators induced by the least squares loss are sensitive to heavy-tailed noise or outliers; in other words, by using the model (2.2), one assumes that every entry of $\mathcal{N}$ obeys the standard Gaussian distribution by default. However, in real-world applications, data may be contaminated by heavy-tailed noise, and even outliers/impulsive noise, which least squares–based estimation may not be capable of due to its non-robustness. Here by robustness/non-robustness, we mean that the model is resistant/not resistant to heavy-tailed noise or outliers. A canonical approach to address this problem is to model noise by using heavy-tailed distributions. A typical choice to this purpose is to use Cauchy distributions, whose probability density function can be expressed as

$$P_{\text{Cauchy}}(t) \propto \frac{1}{1 + (t - c)^2/\delta^2},$$  \hspace{1cm} (2.3)

where $\delta > 0$ is the scale parameter and $c$ is the location parameter. By assuming the symmetry of the noise, we let $c = 0$ in the above function.
To deal with (2.1) in the presence of Cauchy noise (or even other heavy-tailed noise or outliers), we propose the following optimization model, with its derivation from the maximum a posterior (MAP) estimation viewpoint being given later:

$$\min_{\Phi, U} \Phi_\delta(A - [\sigma; U]) := \frac{\delta^2}{2} \sum_{i_1, \ldots, i_d = 1}^{n_1, \ldots, n_d} \log \left(1 + \left([\sigma; U]_{i_1 \ldots i_d} - A_{i_1 \ldots i_d}\right)^2 / \delta^2\right)$$

s.t. $u_{j,i} \in \text{st}(n_j, 1)$, $1 \leq j \leq d - t$, $1 \leq i \leq R$,

$$U_j \in \text{st}(n_j, R), d - t + 1 \leq j \leq d. \quad (2.4)$$

Comparing (2.4) with (2.2), the difference is that the least squares loss is replaced by the statistically motivated loss function

$$\phi_\delta(t) := \frac{\delta^2}{2} \log \left(1 + t^2 / \delta^2\right). \quad (2.5)$$

$\phi_\delta(\cdot)$ is called the Cauchy loss. From the plots in the left panel of Fig. 2.4, it can be observed that, comparing with the squares loss, Cauchy losses can suppress the effect of large errors by tuning the $\sigma$ values, which intuitively explains the robustness of Cauchy loss–based optimization model (2.4). Due to their effectiveness in tackling heavy-tailed noise, in recent years, various Cauchy loss–based models have been proposed in different areas to deal with robust estimation problems. For instance, in [12, 27, 39, 43], a Cauchy loss–based model was proposed to remove Cauchy noise in images; in [17, 36], the Cauchy loss is utilized in robust matrix factorization and subspace clustering problem; in [19], it is shown that the Cauchy loss can be used to address robust face recognition problems; in [54], a Cauchy loss–based regularized model is proposed to deal with robust tensor recovery problem.

Here, we discuss some properties of the proposed model (2.4) from the robust statistics viewpoint, which shows (2.4) is not only resistant to Cauchy noise, but may also be resistant to other heavy-tailed noise or outliers. Firstly, we observe that

$$\lim_{|t| \to +\infty} \phi_\delta'(t) = \lim_{|t| \to +\infty} \frac{t}{1 + t^2 / \delta^2} = 0. \quad (2.6)$$

Such a property is called the redescending property in robust statistics [26], and the minimizer of (2.4) is called a redescending M-estimator. It is known that the redescending M-estimator is robust to heavy-tailed noise and outliers [26]. As a comparison, the derivative of the least squares loss $\ell_2(t) = t^2 / 2$ is $t$, whose limit is infinity, which does not have the redescending property. Other loss functions admitting the redescending property include the Welsch loss [13, 14, 21], the Tukey loss [3], the German loss [15], and so on.

Secondly, the parameter $\delta$ in (2.5) controls the robustness of the model (2.4). From (2.6), we see that the smaller $\delta$ is, the faster $\phi_\delta'(t)$ converges to zero. We plot $\phi_\delta(t)$ with different $\delta$ in the right panel of Fig. 1. On the other hand, taking Taylor expansion of $\phi_\delta(t)$ at 0 yields $\phi_\delta(t) = t^2 / 2 + o(t^2 / \delta^2)$, which shows that $\phi_\delta(t) \approx t^2 / 2$ as $\delta \to \infty$. These observations imply that a small $\delta$ can enhance the robustness of (2.4).
This also reminds us that (2.4) is also resistant to Gaussian noise by simply setting a large enough $\delta$. We also plot $\phi_3(t)$ with different $\delta$ in the left panel of Fig. 1.

**Derivation of (2.4) from MAP estimation** We derive (2.4) from MAP estimation by assuming that $\mathcal{N}$ obeys the Cauchy distribution (2.3). To this end, denote respectively the indicator function $1_C(\cdot)$ and the characteristic function $\iota_C(\cdot)$ of a closed set $C$ as follows

$$1_C(x) = 1, \text{ if } x \in C; \quad 1_C(x) = 0, \text{ if } x \notin C,$$

$$\iota_C(x) = 0, \text{ if } x \in C; \quad \iota_C(x) = +\infty, \text{ if } x \notin C.$$

From the constraints on $u_{j,i}$ and $U_j$, it is natural to impose a uniform prior belief distributional assumption on $u_{j,i}, 1 \leq j \leq d - t, 1 \leq i \leq R$, and $U_j, d - t + 1 \leq j \leq R$ as follows

$$P(\sigma; U) \propto \prod_{j=1}^{d-t} \prod_{i=1}^R 1_{st(n_j,i)}(u_{j,i}) \cdot \prod_{j=d-t+1}^d 1_{st(n_j,R)}(U_j). \quad (2.7)$$

On the other hand, in the presence of Cauchy noise, the probability of the observed data tensor $A$ conditioned on $\sigma; U$ is given by

$$P\left(A_{i_1 \ldots i_d} \mid \sigma; U\right) \propto \frac{1}{1 + \left(\|\sigma; U\|_{i_1 \ldots i_d} - A_{i_1 \ldots i_d}\right)^2 / \delta^2}. \quad (2.8)$$
where $1 \leq i_j \leq n_j$, $1 \leq j \leq d$. With (2.7) and (2.8) at hand, using Bayes’s rule, the MAP estimation is given by

$$\{\sigma^*, U^*\} = \arg \max P\left(\sigma; U \mid A\right)$$

$$= \arg \max \frac{P(A \mid \sigma; U) \cdot P(\sigma; U)}{P(A)}$$

$$= \arg \max \prod_{i_1=1}^{n_1} \ldots \prod_{i_d=1}^{n_d} P\left(\mathcal{A}_{i_1 \ldots i_d} \mid [\sigma; U]_{i_1 \ldots i_d}\right) \cdot P(\sigma; U)$$

$$t \leftarrow \log(\tau)$$

$$= \arg \min \sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} \log \left(1 + \left(\frac{\|\sigma; U\|_{i_1 \ldots i_d} - \mathcal{A}_{i_1 \ldots i_d}}{\delta^2}\right)^2\right)$$

$$- \sum_{j=1}^{d-t} \sum_{i=1}^{R} \log \left(1_{st(n_j, 1)}(u_{j, i})\right) - \sum_{j=d-t+1}^{d} \log \left(1_{st(n_j, R)}(U_j)\right)$$

$$= \arg \min \sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} \log \left(1 + \left(\frac{\|\sigma; U\|_{i_1 \ldots i_d} - \mathcal{A}_{i_1 \ldots i_d}}{\delta^2}\right)^2\right)$$

$$+ \sum_{j=1}^{d-t} \sum_{i=1}^{R} t_{st(n_j, 1)}(u_{j, i}) + \sum_{j=d-t+1}^{d} t_{st(n_j, R)}(U_j), \quad (2.9)$$

where in the last equality, we have defined $\log(0) = -\infty$. Note that (2.9) is exactly the same as (2.4). This gives the explanation of (2.4) from the MAP estimation viewpoint.

**Remark 2.1** We discuss several differences between our model (2.4) and some existing robust tensor models. In recent years, robust techniques have been incorporated into tensor decomposition/approximation/recovery/completion/PCA problems, where the $L_1$ loss function, namely, $\ell_1(t) = |t|$, is frequently employed to deliver robustness. In general, such kind of models can be formulated as [16]

$$\min_{\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}} \|L(\mathcal{X}) - b\|_1 + \lambda R(\mathcal{X}), \quad (2.10)$$

where $L$ is a linear operator, and $b$ has the same size as $L(\mathcal{X})$; $R(\mathcal{X})$ denotes a certain regularizer that controls the low-rankness of $\mathcal{X}$, such as the sum of nuclear norms of unfolding matrices of $\mathcal{X}$ [47], and $\lambda > 0$ is the regularization parameter. A special case of (2.10) is the robust tensor PCA, in which $L$ is the identity operator and $b$ denotes the observed tensor [16]. It is known that $L_1$ loss is more suitable for Laplacian noise; on the other hand, one sees that the derivative of $|t|$ does not tend to zero as $|t| \to +\infty$, meaning that it does not admit the redescending property, while it was pointed out in [38] that the $L_1$ estimator might behave as bad as the $\ell_2(t)$ estimator in some cases. Comparing with the resulting tensor, (2.10) yields a full tensor of size $n_1 \times \cdots \times n_d$, while ours is compressed into a set of factor matrices, which takes much less storage. Moreover, the orthornormality structure on some factor matrices is more suitable for certain applications [8–10, 44, 46, 49].
In [1], a robust tensor CP decomposition model has been considered. The differences are that the noise there is required to be sparse, and all the factor matrices are assumed to be columnwisely orthogonal, which are stringent. By using outlier detection techniques, [41] proposed a robust Tucker model. However, the underlying model cannot be clearly formulated as an optimization problem, and the tensor model is different from ours. By using variational inference and Kullback-Leibler divergence, [6] devised a robust algorithm to find CP approximation with orthonormal factors, where the model and the solution method are quite different from ours. In particular, the authors pointed out that their algorithm boils down to the alternating least squares [48] in the absence of outliers. In a recent survey [22], various statistically motivated loss functions are incorporated into tensor CPD, in which the Huber’s loss is considered. As Huber’s loss can be regarded as a smoothed $\ell_1$ loss, it does not admit the redescending property as well. The orthonormality is not taken into account in [22]. Note that the idea of employing Cauchy loss has been considered in the authors’ earlier work [54]. Comparing with (2.4), the resulting tensor in [54] is a full tensor and also does not take into account the orthonormality, and the solution method is also different.

The remaining problem is how to solve (2.4) efficiently. For this purpose, several properties concerning the Cauchy loss for designing and analyzing the solution method are first introduced in the following subsection.

### 2.1 Quantitative properties concerning $\phi_\delta(\cdot)$

First, we introduce the so-called half-quadratic (HQ) property of $\phi_\delta(\cdot)$, which turns the function into a weighted least squares problem and is crucial for designing the algorithm. Such a property of the Cauchy loss has appeared in the literature (see, e.g., [17, 19]) in which the verification is based on the utilization of conjugate functions. Here for completeness, we present a very direct and concise verification. Recall that we have defined $\log(0) = -\infty$.

**Lemma 2.1** (Half-quadratic property) Given $|t| < +\infty$, it holds that

$$\phi_\delta(t) = \min_{\omega \geq 0} \frac{\omega}{2} t^2 + \frac{\delta^2}{2} \varrho(\omega),$$

where $\varrho(\omega) = \omega - \log(\omega) - 1$. Moreover, the minimizer of (2.11) is given by

$$\omega^* = \frac{\delta^2}{\delta^2 + t^2}.$$  \hspace{1cm} (2.12)

**Proof** First we verify that (2.12) is a minimizer of the right hand-side of (2.11). Denote $g(\omega) := \omega t^2 / 2 + \delta^2 \varrho(\omega) / 2$. As $\varrho(\cdot)$ is convex, it suffices to show that $\omega^*$ in (2.12) is a stationary point of $\inf_{\omega \geq 0} g(\omega)$. Since $|t| < +\infty$, we see that the minimizer
of $\inf_{\omega \geq 0} g(\omega)$ cannot occur at $\omega = 0$. Thus, any stationary point of $\inf_{\omega \geq 0} g(\omega)$ meets

$$g'(\omega) = 0 \Leftrightarrow t^2 + \delta^2 - \frac{\delta^2}{\omega} = 0,$$

and so $\omega = (1 + t^2/\delta^2)^{-1}$, which is exactly (2.12). Inserting this expression into (2.11), we get

$$2g(t) = \frac{\delta^2}{\delta^2 + t^2} (t^2 + \delta^2) + \delta^2 \log(1 + t^2/\delta^2) - \delta^2 = \delta^2 \log(1 + t^2/\delta^2),$$

boiling down to the expression of $\phi_\delta(t)$. The proof is completed. \qed

Note that the HQ property has a clear indication on robustness: Take $t = A_{i_1 \ldots i_d} - [\sigma; U]_{i_1 \ldots i_d}$ in Lemma 2.1 as the noise; we see that the larger the magnitude of $t$, the smaller the weight $\omega$ it yields, and so the corresponding term $\phi_\delta(t)$ is less important in the objective $\Phi_\delta(\cdot)$ in (2.4).

Recalling that $\phi_\delta(t) = \frac{\delta^2 t}{\delta^2 + t^2}$, we have:

**Proposition 2.2** (Lipschitz gradient) For any $t_1, t_2 \in \mathbb{R}$ and $\delta > 0$, it holds that

$$\left| \frac{\delta^2 t_1}{\delta^2 + t_1^2} - \frac{\delta^2 t_2}{\delta^2 + t_2^2} \right| \leq |t_1 - t_2|.$$

**Proof** By the mean value theorem, it suffices to show that $|\phi_\delta''(t)| \leq 1$. In fact,

$$|\phi_\delta''(t)| = \left| \frac{\delta^2 (\delta^2 - t^2)}{(\delta^2 + t^2)^2} \right| \leq \frac{\delta^2}{\delta^2 + t^2} \leq 1,$$

and the result follows. \qed

The following Lipschitz-like inequality is important for convergence analysis.

**Proposition 2.3** (Lipschitz-like inequality) Let $t_1, t_2 \in \mathbb{R}$ be arbitrary, and let $\delta > 0$. Then it holds that

$$|e| := \left| \delta^2 t_1 \left( \frac{1}{\delta^2 + t_1^2} - \frac{1}{\delta^2 + t_2^2} \right) \right| \leq |t_1 - t_2|.$$

**Proof** It is clear that

$$|e| = \left| \delta^2 t_1 \frac{(t_1 + t_2)(t_1 - t_2)}{(\delta^2 + t_1^2)(\delta^2 + t_2^2)} \right| \leq \delta^2 |t_1| \frac{|t_1| + |t_2|}{(\delta^2 + t_1^2)(\delta^2 + t_2^2)} \cdot |t_1 - t_2|.$$

To prove the above relation, it suffices to show the coefficient of $|t_1 - t_2|$ is not greater than 1, i.e.,

$$\varphi(t_1, t_2) := (\delta^2 + t_1^2)(\delta^2 + t_2^2) - \delta^2 |t_1|(|t_1| + |t_2|) \geq 0.$$
In fact,
\[ \varphi(t_1, t_2) = \delta^4 + \delta^2 t_2^2 + |t_1 t_2|(|t_1 t_2| - \delta^2) \geq \delta^4 + \delta^2 t_2^2 - \frac{\delta^4}{4} \geq 0. \]
Therefore, \( |e| \leq |t_1 - t_2| \), as desired.

\[ \square \]

3 HQ-ADMM

By using Lemma 2.1, we equivalently rewrite the objective function \( \Phi_\delta(\cdot) \) of (2.4) in what follows. Specifically, since \( \Phi_\delta(\cdot) \) is the sum of \( \phi_\delta(\cdot) \) functions, taking \( t = A_{i_1 \cdots i_d} - [\sigma; U]_{i_1 \cdots i_d} \) in Lemma 2.1, we have
\[
\Phi_\delta(A - [\sigma; U]) = \frac{1}{2} \min_{W_{i_1 \cdots i_d} \geq 0} \sum_{i_1=1, \ldots, i_d=1}^{n_1 \cdots n_d} \left[ \mathcal{W}_{i_1 \cdots i_d} (A_{i_1 \cdots i_d} - [\sigma; U]_{i_1 \cdots i_d})^2 + \delta^2 q(\mathcal{W}_{i_1 \cdots i_d}) \right].
\]
where we denote \( \mathcal{W} = (\mathcal{W}_{i_1 \cdots i_d}) \in \mathbb{R}^{n_1 \times \cdots \times n_d} \) as a tensor variable. From Lemma 2.1, we see that the optimizer is
\[
\mathcal{W}_{i_1 \cdots i_d} = \delta^2 \left( 1 + (\|\sigma; U\|_{i_1 \cdots i_d} - A_{i_1 \cdots i_d})^2 / \delta^2 \right)^{-1}.
\]
As explained in the paragraph below Lemma 2.1, \( \mathcal{W} \) can be interpreted as weights to the problem. From the expression of \( \mathcal{W} \), we see that the larger the noise is, the smaller the weight gives to the problem. Such a mechanism helps mitigate heavy-tailed noise or outliers.

In view of (3.1), a straightforward idea to solve (2.4) (with the objective replaced by (3.1)) is to employing an alternating minimization method (AMM) by iteratively updating \( \sigma, U, \) and \( \mathcal{W} \). In fact, applying AMM to solve Cauchy loss–based problems have been considered in the literature (see, e.g., [17, 19]). However, for our problem, this leads to that the subproblems related to \( U_j \) do not have closed-form solutions. [36] also applied AMM to solve Cauchy loss–based problem; however, as their proposed model is unconstrained and the objective function is smooth, AMM yields closed-form solutions to each subproblem. [43] incorporated Cauchy loss into models for image processing. However, the problem is convexified by imposing a quadratic term, which results in that the Cauchy loss–related subproblem admits a unique solution that can be analytically solved by solving a cubic equation. If the subproblem is nonconvex, then numerical methods have to be applied to solving the Cauchy loss–related subproblem, as pointed out in [43], which might result in inefficiency. For other Cauchy loss–based image processing problems, [12, 27, 39] proposed to use the conventional alternating direction method of multipliers (ADMM) directly. However, without noticing the HQ property, in the ADMM, solving the Cauchy loss–related subproblem also does not admit a closed-form solution. As a result, solving such a subproblem still requires an iterative method. [54] used a linearization technique, which ignored the HQ property.
In view of the above limitations in dealing with Cauchy loss–based problems, in this section, by combining the HQ property and the ADMM framework, we develop the HQ-ADMM to solve our model (2.4). The advantage of HQ-ADMM is that all the subproblems involved in the algorithm admit closed-form solutions. In what follows, we derive our method step by step.

Note that (3.1) is quadratic with respect to each $U_j$, leading to the following formulation

$$
\Phi_\delta(A-\tau; U) = \frac{1}{2} \min_{\tau \geq 0} \| \sqrt{\mathcal{W}} \otimes (A-\tau; U) \|^2_F + \frac{\delta^2}{2} \sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} \phi(\mathcal{W}_{i_1 \ldots i_d}),
$$

where $\sqrt{\mathcal{W}} = (\sqrt{\mathcal{W}_{i_1 \ldots i_d}}) \in \mathbb{R}^{n_1 \times \ldots \times n_d}$ and “$\otimes$” denotes the Hadamard product. With this expression at hand, by introducing a slack variable $\mathcal{T} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$, we rewrite (2.4) as

$$
\min_{\sigma, U, \mathcal{T}, \mathcal{W}} \Phi_\delta(A - \mathcal{T}) = \frac{1}{2} \| \sqrt{\mathcal{W}} \otimes (A - \mathcal{T}) \|^2_F + \frac{\delta^2}{2} \sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} \phi(\mathcal{W}_{i_1 \ldots i_d})
$$

s.t. $\mathcal{T} = [\sigma; U]$, $\mathcal{W} \geq 0$, $\mathcal{u}_{j,i}^T \mathcal{u}_{j,i} = 1, 1 \leq j \leq d - t, 1 \leq i \leq R$, $U_j^T U_j = I, d - t + 1 \leq j \leq d$. (3.2)

By introducing a Lagrangian multiplier $\mathcal{Y} \in \mathbb{R}^{n_1 \times \ldots \times n_d}$, the augmented Lagrangian function of (3.2) is given by

$$
L_{\tau}(\sigma, U, \mathcal{T}, \mathcal{Y}, \mathcal{W}) := \frac{1}{2} \| \sqrt{\mathcal{W}} \otimes (A - \mathcal{T}) \|^2_F + \frac{\delta^2}{2} \sum_{i_1=1}^{n_1} \ldots \sum_{i_d=1}^{n_d} \phi(\mathcal{W}_{i_1 \ldots i_d}) - \frac{\tau}{2} \| [\sigma; U] - \mathcal{T} \|^2_F,
$$

where $\tau > 0$. In what follows, for notational convenience we denote $(\mathcal{Y} + \tau \mathcal{T}) \otimes_{j \neq j} d_{i,j} \mathcal{u}_{j,i} \in \mathbb{R}^{n_j}$ as the gradient of $(\mathcal{Y} + \tau \mathcal{T}) \otimes_{j=1}^{d} \mathcal{u}_{j,i}$ with respect to $\mathcal{u}_{j,i}$. Then, the last two terms of (3.3) can be rewritten as

$$
- \langle \mathcal{Y}, [\sigma; U] - \mathcal{T} \rangle + \frac{\tau}{2} \| [\sigma; U] - \mathcal{T} \|^2_F
$$

(3.4)
to $U_j^\top U_j = I$, where $\Lambda_j$’s are symmetric matrices. Denote

$$
\hat{L}_r(\sigma, U, \mathcal{T}, \mathcal{Y}, \mathcal{W}) := L_r(\sigma, U, \mathcal{T}, \mathcal{Y}, \mathcal{W})
+ \sum_{j,i=1}^{d-t} \eta_{j,i} \left(u_{ij}u_{ij}^\top - 1\right) + \sum_{j=d-t+1}^{d} \left(\Lambda_j, U_j^\top U_j - I\right). 
$$

Thus, taking derivative of $\hat{L}(\cdot)$ with respect to each $u_{j,i}, 1 \leq j \leq d-t, 1 \leq i \leq R$ and noticing (3.4) yields

$$
\sigma_i(\mathcal{Y} + \tau \mathcal{T}) \otimes_{l\neq j}^d u_{i,l} = \eta_{j,i} u_{j,i}, 1 \leq j \leq d-t, 1 \leq i \leq R. \tag{3.6}
$$

Since $u_{j,i}$’s are normalized, we get $\eta_{j,i} = \sigma_i \mathcal{Y} + \tau \mathcal{T}$, $\otimes_{l=1}^d u_{i,l}$. On the other hand, noticing the representation (3.4), taking derivative of $\hat{L}(\cdot)$ with respect to $\sigma$ gives that $\sigma_i = (\mathcal{Y} + \tau \mathcal{T}, \otimes_{l=1}^d u_{i,l})/\tau$, which together with the expression of $\eta_{j,i}$ gives $\eta_{j,i} = \sigma_i^2 \tau$; therefore, (3.6) is in fact as follow

$$
(\mathcal{Y} + \tau \mathcal{T}) \otimes_{l\neq j}^d u_{i,l} = \sigma_i \tau u_{j,i}, 1 \leq j \leq d-t, 1 \leq i \leq R. \tag{3.7}
$$

Next, taking derivative with respect to $u_{j,i}, d-t + 1 \leq j \leq d, 1 \leq i \leq R$ and noticing (3.4) gives

$$
\sigma_i(\mathcal{Y} + \tau \mathcal{T}) \otimes_{l\neq j}^d u_{i,l} = \sum_{r=1}^{R} (\Lambda_j)_{i,r} u_{j,r}, 1 \leq j \leq d-t, 1 \leq i \leq R. \tag{3.8}
$$

Denote $\mathcal{E} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ as the all-one tensor; taking derivative with respect to $\mathcal{T}$ and rearranging terms yields

$$
\mathcal{W} \otimes (\mathcal{T} - \mathcal{A}) + \mathcal{Y} - \tau (\mathcal{Y} + \mathcal{T}) = 0 
\Leftrightarrow (\mathcal{W} + \tau \mathcal{E} \otimes \mathcal{T} = \mathcal{W} \otimes \mathcal{A} - \mathcal{Y} + \tau \mathcal{Y} + \mathcal{U}]. \tag{3.9}
$$

As a result, taking (3.7), (3.8), (3.9), and Lemma 2.1 into account, any stationary point $\{\sigma, U, \mathcal{T}, \mathcal{Y}, \mathcal{W}\}$ satisfies the following system

$$
\begin{align*}
(\mathcal{Y} + \tau \mathcal{T}) \otimes_{l\neq j}^d u_{i,l} &= \sigma_i \tau u_{j,i}, & 1 \leq j \leq d-1, 1 \leq i \leq R, \\
u_{j,i}^\top u_{j,i} &= 1, & 1 \leq j \leq d-1, 1 \leq i \leq R, \\
\sigma_i(\mathcal{Y} + \tau \mathcal{T}) \otimes_{l\neq j}^d u_{i,l} &= \sum_{r=1}^{R} (\Lambda_j)_{i,r} u_{j,r}, & 1 \leq j \leq d-1, 1 \leq i \leq R, \\
U_j^\top U_j &= I, & 1 \leq j \leq R, \\
(\mathcal{W} + \tau \mathcal{E} \otimes \mathcal{T} &= \mathcal{W} \otimes \mathcal{A} - \mathcal{Y} + \tau \mathcal{Y} + \mathcal{U}]. \\
[\mathcal{Y} + \tau \mathcal{T}] &= \mathcal{T}.
\end{align*}
\tag{3.10}
$$

**HQ-ADMM framework** Combining the HQ property and the ADMM, the HQ-ADMM computes the following subproblems at each iterate

$$
\begin{align*}
U_j^{k+1} &\in \arg \min_{\|u_{j,l}\|_2^2, 1 \leq l \leq R} L_r(\sigma^k, U_1^{k+1}, \ldots, U_j^{k+1}, \ldots, U_d^{k+1}, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k), & 1 \leq j \leq d-t, \\
U_j^{k+1} &\in \arg \min_{\|u_{j,l}\|_2^2, 1 \leq l \leq R} L_r(\sigma^k, U_1^{k+1}, \ldots, U_{j-1}^{k+1}, U_j^{k+1}, \ldots, U_d^{k+1}, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k), & 1 \leq j \leq d-t, \\
\mathcal{T}^{k+1} &\in \arg \min_{\mathcal{T} \in \mathcal{T}} L_r(\sigma^k, U_1^{k+1}, \ldots, U_{j-1}^{k+1}, U_j^{k+1}, \ldots, U_d^{k+1}, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k), & d-t + 1 \leq j \leq d,
\end{align*}
$$

$$
\begin{align*}
\mathcal{Y}^{k+1} &= \mathcal{T}^k - \tau (\mathcal{Y}^k) - \mathcal{T}^k, \\
\sigma^{k+1} &= \arg \min_{\sigma \in \mathcal{Y}} L_r(\sigma, U_1^{k+1}, \ldots, U_d^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^k), \\
\mathcal{W}^{k+1} &= \arg \min_{\mathcal{W}} L_r(\sigma^{k+1}, U_1^{k+1}, \ldots, U_d^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^k).
\end{align*}
$$
Comparing with the standard ADMM framework, HQ-ADMM involves an additional subproblem to update the weights $\mathcal{W}$. In what follows, we present how to solve each subproblem.

$U_j$-subproblems For notational convenience, let

$$v_{j,i}^{k+1} := (\mathcal{Y}^k + \tau T^k)u_{j,i}^{k+1} \otimes \cdots \otimes u_{j-1,i}^{k+1} \otimes u_{j+1,i}^k \otimes \cdots \otimes u_{d,i}^k$$

represent the gradient of $\left(\mathcal{Y}^k + \tau T^k, \otimes_i^d u_{j,i}^k\right)$ with respect to $u_{j,i}$ at the point $(u_{j,i}^{k+1}, \ldots, u_{j-1,i}^{k+1}, u_{j+1,i}^k, \ldots, u_{d,i}^k)$. Denote $V_j^{k+1} := [v_{1,i}^{k+1}, \ldots, v_{j+1,i}^{k+1}]$ and $\sigma_i^k$. When $1 \leq j \leq d - t$, from the definition of $L_\tau()$, $v_{j,i}$, and noticing (3.3) and (3.4), each $u_{j,i}$ can be updated as follows

$$u_{j,i}^{k+1} = \arg \min_{\|u_{j,i}\|=1} -\sigma_i^k\left(v_{j,i}^{k+1}, u_{j,i}\right) \iff u_{j,i}^{k+1} = v_{j,i}^{k+1}/\|v_{j,i}^{k+1}\|, \ 1 \leq i \leq R. $$

However, for theoretical convergence purpose we compute the following instead

$$u_{j,i}^{k+1} = \tilde{v}_{j,i}^{k+1}/\|\tilde{v}_{j,i}^{k+1}\|, \text{ where } \tilde{v}_{j,i}^{k+1} = \sigma_i^k v_{j,i}^{k+1} + \alpha u_{j,i}^k, \ 1 \leq i \leq R; \quad (3.11)$$

here $\alpha > 0$ is an arbitrary constant.

When $1 \leq j \leq d$, from the definition of $L_\tau()$, $v_{j,i}^{k+1}$, $V_j^{k+1}$ and recalling (3.3) and (3.4), it follows

$$U_j^{k+1} = \arg \min_{U_j^i U_j=I} \sum_{i=1}^R \langle \sigma_i v_{j,i}^{k+1}, u_{j,i}\rangle = \arg \max_{U_j^i U_j=I} \left(V_j^{k+1} \cdot \text{diag}(\sigma^k), U_j\right),$$

where $\text{diag}(\sigma) = [\sigma_1, \ldots, \sigma_R] \in \mathbb{R}^{R \times R}$ is a diagonal matrix. Similar to (3.11), we in fact compute the following problem instead

$$U_j^{k+1} = \arg \max_{U_j^i U_j=I} \left(\tilde{V}_j^{k+1}, U_j\right), \text{ where } \tilde{V}_j^{k+1} = V_j^{k+1} \cdot \text{diag}(\sigma^k) + \alpha U_j^k. \quad (3.12)$$

The above problem is to compute the polar decomposition of $\tilde{V}_j^{k+1}$, which admits a closed-form solution. Specifically, assume $\tilde{V}_j^{k+1} = \tilde{P} \tilde{Q} \tilde{Q}^\top$ is the compact SVD of $\tilde{V}_j^{k+1}$, where $P \in \mathbb{R}^{n_j \times R}$, $Q \in \mathbb{R}^{R \times R}$, $P^\top P = I$, $Q^\top Q = Q Q^\top = I$, $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_R)$ with $\lambda_i$ being the singular value of $\tilde{V}_j^{k+1}$. Then, $U_j^{k+1} = P Q \Sigma$. Moreover, letting $H_j^{k+1} := Q \Sigma Q^\top$. Then,

$$\tilde{V}_j^{k+1} = U_j^{k+1} H_j^{k+1}. \quad (3.13)$$

$T$-, $\sigma$- and $\mathcal{W}$-subproblems From (3.9), we have that

$$T_j^{k+1} = \left(\mathcal{W}^k_{i_1 \cdots i_d} A_{i_1 \cdots i_d} - \mathcal{Y}^k_{i_1 \cdots i_d} + \tau \left[\sigma^k; U_j^{k+1}\right]_{i_1 \cdots i_d}\right) / \left(\mathcal{W}^k_{i_1 \cdots i_d} + \tau\right). \quad (3.14)$$

To compute $\sigma_{i}^{k+1}$, from the expression of (3.4) it is easily seen that

$$\sigma_{i}^{k+1} = (\mathcal{Y}^{k+1} + \tau T^{k+1}) \otimes_i^d u_{j,i}^{k+1} / \tau, \ 1 \leq i \leq R. \quad (3.15)$$
To compute $\mathcal{W}^{k+1}$, similar to (3.10) we have
\[
\mathcal{W}^{k+1}_{i_1 \ldots i_d} = \delta^2 \left( \delta^2 + \left( \mathcal{T}^{k+1}_{i_1 \ldots i_d} - \mathcal{A}_{i_1 \ldots i_d} \right)^2 \right)^{-1}.
\] (3.16)

In summary, the HQ-ADMM is described in Algorithm 1, where each subproblem admits a closed-form solution.

**Algorithm 1** HQ-ADMM for solving (2.4).

1. **Require:** $U^0_j = [u^0_{j,i}, \ldots, u^0_{i,R}]$, with $\|u_{j,i}\| = 1, 1 \leq j \leq d - t, 1 \leq i \leq R$;
2. \((U^0)^\top U^0_j = I, d - t + 1 \leq j \leq d; \sigma^0, \mathcal{T}^0, \mathcal{Y}^0, \mathcal{W}^0, \alpha > 0, \tau > 0, \delta > 0.\)
3. for \( k = 0, 1, \ldots, \)
5. Compute $\mathcal{U}^{k+1}_j$ via (3.12), $d - t + 1 \leq j \leq d$
4. Compute $\mathcal{T}^{k+1}$ via (3.14),
5. Compute $\mathcal{Y}^{k+1} = \mathcal{Y}^k - \tau \left( \|\sigma^k; U^{k+1}\| - \mathcal{T}^{k+1} \right)$,
6. Compute $\sigma^{k+1}$ via (3.15),
7. Compute $\mathcal{W}^{k+1}$ via (3.16).
8. end for

**Remark 3.2 1.** The idea of HQ-ADMM can be applied to a more general form of (2.4). Specifically, consider the data-fitting term given by $\Phi_\delta \left( L (\|\sigma; U\|) \right) - b$), where $L$ is a linear operator, and $b$ denotes the observed data of the same size as $L (\|\sigma; U\|)$. When $L$ represents the identity operator and $b$ denotes $\mathcal{A}$, the data-fitting term boils down to the objective of (2.4). When $\Phi_\delta \left( L (\|\sigma; U\|) \right) - b = \Phi_\delta (\Omega \oplus (\|\sigma; U\| - \mathcal{A})), where $\Omega \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is a given $0 - 1$ tensor with $\Omega_{i_1 \ldots i_d} = 1$ if $\mathcal{A}_{i_1 \ldots i_d}$ is being observed while $\Omega_{i_1 \ldots i_d} = 0$ if $\mathcal{A}_{i_1 \ldots i_d}$ missing, it can be used to deal with robust tensor approximation with incomplete data. When $L$ is formed by a set of input data tensors, and each entry of $b$ denotes the output score of the corresponding input data tensor, it is the objective of the (robust) tensor regression problem.

To minimize $\Phi_\delta \left( L (\|\sigma; U\|) \right) - b$, similar to (3.2), one can also formulate the problem as
\[
\min_{\sigma, U, T, \mathcal{W}} \Phi_\delta \left( L (T) - b \right) = \frac{1}{2} \|\sqrt{\mathcal{W}} \oplus (L (T) - b)\|^2_F + \delta^2 \sum_{i=1}^d \mathcal{Q}(w_i) \\
\text{s.t. } T = [\|\sigma; U\|, \mathcal{W} \geq 0, \\
u_{j,i}^\top u_{j,i} = 1, 1 \leq j \leq d - t, 1 \leq i \leq R, \\
u_{j,i}^\top u_{j,i} = I, d - t + 1 \leq j \leq d,
\]

where $w$ is the same size as $b$ defined similar to that in (3.1). The framework of HQ-ADMM then applies as well; the main difference is that solving the $T$-subproblem amounts to solving a linear system of equations. The convergence
analysis and convergence results presented in the next section also apply without many modifications.

2. HQ-ADMM can also be applied to solve other Cauchy loss–based problems such as those studied in [27, 39]. Specifically, for problems of the form

$$\min_x \Phi_\delta(Lx - b) + R(x),$$

where $L$ is a matrix, $b$ is a vector of proper size, one can also convert it to

$$\min_{x, w} \| \sqrt{w} \otimes (Ly - b) \|_F^2 + \delta^2 \sum_{i=1}^p g(w_i) + R(x), \text{ s.t. } x = y,$$

with $w$ defined similar to that in (3.1); an algorithm in the spirit of HQ-ADMM can be applied to solve it.

3. An alternative way to obtain closed-form solutions in ADMM is to use a linearization technique. For example, one can apply a linearized ADMM to solve the original problem (2.4) instead of the equivalent form (3.2), in which one also replaces $[\sigma; U]$ by $T$; however, to solve the $T$-subproblem, i.e., $\min_T \Phi_\delta(T - A) + \langle Y T \rangle + \tau / 2 \| T - [\sigma; U] \|_F^2$, which does not admit a closed-form solution, one linearizes $\Phi_\delta(T - A)$ and then imposes a proximal term. This has some similarities with HQ-ADMM; however, by doing this, one does not fully explore the HQ property of the model, and extra effort has to be paid to find a suitable step-size for this linearized subproblem, which may lead to inefficiency.

4. HQ-ADMM can be categorized as a manifold ADMM [29] since the involved optimization problem is constrained on manifolds.

5. HQ-ADMM can be regarded as an inexact alternating minimization for solving (3.2): One first update the variables $\sigma$, $U$, $T$ by running one step of the standard ADMM, and then update $W$.

6. The inexact augmented Lagrangian method (iALM) can possibly applied to solve (3.2), while the main difference is that the primal variables have to be simultaneously updated when using iALM, which might not admit closed-form solutions.

7. The update of $\sigma$ can also be put right after $U$, which does not affect the convergence. On the other hand, the updates of $W$ and $Y$ only depend on $\sigma$, $U$, and $T$, and are independent of each other. Thus, the update of $W$ can be put before $Y$.

4 Convergence of HQ-ADMM

Convergence of nonconvex ADMM and splitting type methods was established in [23, 30–32, 52]. This section presents the convergence of HQ-ADMM, following a similar way of [31]; however, some properties of the Cauchy loss are employed to overcome certain obstacles that only appear in the analysis of HQ-ADMM. Throughout this section, to simplify the notations, we denote

$$\Delta_{U_j}^{k+1, k} := U_j^{k+1} - U_j^k.$$
The definitions of $\Delta^{k+1,k}_{\mathcal{T}}$, $\Delta^{k+1,k}_{\mathcal{Y}}$, and $\Delta^{k+1,k}_{\mathcal{W}}$ are analogous. In addition, we define the following proximal augmented Lagrangian function
\[
\tilde{L}_{\tau}(\sigma, U, \mathcal{T}, \mathcal{Y}, \mathcal{W}, \mathcal{T}') := L_{\tau}(\sigma, U, \mathcal{T}, \mathcal{Y}, \mathcal{W}) + \frac{2}{\tau} \| \mathcal{T} - \mathcal{T}' \|_{F}^{2},
\]
which is needed to study the diminishing property of the terms $\| \Delta_{U_{j}}^{k+1,k} \|_{F}$ and $\| \Delta^{k+1,k}_{\mathcal{T}} \|_{F}$. For convenience, we also denote
\[
\tilde{L}_{\tau}^{k+1,k} := \tilde{L}_{\tau}(\sigma^{k+1}, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^{k+1}, \mathcal{T}^{k}).
\]

We present the first result, showing that the sequence generated by the algorithm is bounded, and every limit point is a stationary point.

**Theorem 4.1** (Subsequential convergence) Let $\{\sigma^{k}, U^{k}, \mathcal{T}^{k}, \mathcal{Y}^{k}, \mathcal{W}^{k}\}$ be generated by Algorithm 1 with $\tau \geq \sqrt{10}$ and $\alpha > 0$. Then
1. $\{\sigma^{k}, U^{k}, \mathcal{T}^{k}, \mathcal{Y}^{k}, \mathcal{W}^{k}\}$ is bounded;
2. the sequence $\{\tilde{L}_{\tau}^{k+1,k}\}$ defined in (4.1) is bounded, nonincreasing and convergent;
3. it holds that
\[
\sum_{k=1}^{\infty} \left( \sum_{j=1}^{d} \left\| \Delta_{U_{j}}^{k+1,k} \|_{F}^{2} + \| \Delta^{k+1,k}_{\mathcal{T}} \|_{F}^{2} \right\| < +\infty, \tag{4.2}
\]
and
\[
\| \Delta_{\sigma}^{k+1,k} \|_{F} \to 0, \quad \| \Delta_{\mathcal{Y}}^{k+1,k} \|_{F} \to 0, \quad \| \sigma^{k}; U^{k} - \mathcal{T}^{k} \|_{F} \to 0. \tag{4.3}
\]
Moreover, every limit point $\{\sigma^{*}, U^{*}, \mathcal{T}^{*}, \mathcal{Y}^{*}, \mathcal{W}^{*}\}$ satisfies the optimality condition (3.10). In particular, $\{\sigma^{*}, U^{*}\}$ is also a stationary point of the original problem (2.4).

Next, based on the Kurdyka-Łojasiewicz (KL) property [4] which is widely used for proving the global convergence of nonconvex algorithms, we have that the whole sequence converges to a single limit point.

**Theorem 4.2** (Global convergence) Under the setting of Theorem 4.1, the whole sequence of $\{U^{k}, \mathcal{T}^{k}\}$ converges to a single limit point, i.e.,
\[
\lim_{k \to \infty} U^{k}_{j} = U^{*}_{j}, \quad 1 \leq j \leq d, \quad \lim_{k \to \infty} \mathcal{T}^{k} = \mathcal{T}^{*}.
\]
linearly [37]. A key to establish the linear convergence rate is to show that the exponent of the Łojasiewicz inequality is 1/2. To estimate this exponent, besides the methods studied in [24, 25, 37, 55], several calculus rules investigated in [33, 56] are also nice tools. Currently, most of the aforementioned problems are least squares based; how to investigate the Łojasiewicz exponent for the Cauchy loss–based problem (2.4), and how to study the convergence rate of HQ-ADMM still need further research.

5 Numerical experiments

We evaluate the robustness of model (2.4) solved by HQ-ADMM in this section using synthetic and real data. The least squares–based model (2.2) is used as a comparison. (2.2) is solved by the alternating least squares (ALS) method. All the computations are conducted on an Intel i7-7770 CPU desktop computer with 32 GB of RAM. The supporting software is MATLAB R2015b. The MATLAB package Tensorlab [50] is employed for tensor operations. The MATLAB code of HQ-ADMM is available at https://github.com/yuningyang19/hqadmm

The stopping criterion is that max{\|U_j^{k+1} - U_j^k\|_F / \|U_j^k\|_F} \leq 10^{-6}, or |\phi_\delta(A - \|\sigma^{k+1}; U^{k+1}\|) - \phi_\delta(A - \|\sigma^k; U^k\|)\| \leq 10^{-6}, or k \geq 2000. The parameter \alpha in HQ-ADMM is set to 10^{-8}, \tau \in \{0.7, 1\}; \delta = 0.05.

Synthetic data We consider randomly generated tensors contaminated by different kinds of noises listed in the following:

- \mathcal{A} = \mathcal{A}_0 / \|\mathcal{A}_0\|_F + \beta \cdot \mathcal{N}/\|\mathcal{N}\|_F, where \mathcal{A}_0 is the ground truth tensor specified later, and \mathcal{N} denotes the Cauchy noise, with scale parameter \delta = 0.05. \beta = 0.5;
- \mathcal{A} = \mathcal{A}_0 / \|\mathcal{A}_0\|_F + \mathcal{O}. Here, \mathcal{O} denotes sparse outliers, with sparsity 0.1, i.e., 10% of the entries of \mathcal{A}_0 are contaminated by outliers. Outliers are drawn uniformly from [0, 10];
- \mathcal{A} = \mathcal{A}_0 / \|\mathcal{A}_0\|_F + \beta \cdot \mathcal{N}/\|\mathcal{N}\|_F, where \mathcal{N} denotes Gaussian noise, with \beta = 0.1.

The ground truth tensor \mathcal{A}_0 = \sigma_1 \bigotimes_{j=1}^d u_j, where U_j are randomly drawn from a uniformly distribution in [-1, 1]. U_j, d-t+1 \leq j \leq d, are then made to be columnwisely orthonormal, while the remaining U_j’s are columnwisely normalized. \sigma_j are drawn from Gaussian distribution. We set d = 3, 4, 5, \cdots, n_d, and R = 5 in all the experiments in this part. The initializers for HQ-ADMM and ALS are randomly generated. The reported results are averaged over 50 instances for each case.

Comparisons of HQ-ADMM for solving (2.4) and ALS for solving (2.2) with Cauchy noise are reported in Table 1, where err. = \|\mathcal{A}_0\| / \|\mathcal{A}_0\|_F - \mathcal{A}^* / \|\mathcal{A}^*\|_F, with \mathcal{A}^* = \|\sigma^*; U^*\| the tensor generated by the algorithm. “iter.” denotes the number of iterates, and “time” stands for the CPU time consumed by the algorithm. From the “err.” columns, we see that in all cases, HQ-ADMM performs much better than ALS; in particular, “err.” of HQ-ADMM is smaller than 0.1 in almost all cases, which confirms that the proposed model and algorithm are consistent with Cauchy noise.
Table 1 Comparison of HQ-ADMM for (2.4) and ALS for (2.2) when the ground truth tensor is contaminated by Cauchy noise

| n  | (d, t) | HQ-ADMM for (2.4) | ALS for (2.2) |
|----|--------|------------------|---------------|
|    |        | err. iter. time   | err. iter. time |
| 10 | (3, 1) | 5.57E-02 395 0.16 | 4.29E-01 149 0.04 |
| 20 | (3, 1) | 4.66E-02 315 0.21 | 4.20E-01 147 0.05 |
| 50 | (3, 1) | 4.30E-02 45 0.09 | 4.33E-01 309 0.27 |
| 80 | (3, 1) | 3.05E-02 71 0.77 | 4.31E-01 190 1.16 |
| 90 | (3, 1) | 3.04E-02 47 0.76 | 4.29E-01 152 1.28 |
| 100| (3, 1) | 3.21E-02 86 1.62 | 4.41E-01 210 1.82 |
| 10 | (3, 2) | 5.25E-02 453 0.19 | 3.84E-01 33 0.01 |
| 20 | (3, 2) | 2.93E-02 137 0.10 | 4.12E-01 17 0.01 |
| 60 | (3, 2) | 2.25E-02 200 1.03 | 4.42E-01 11 0.04 |
| 80 | (3, 2) | 2.20E-02 58 0.60 | 4.18E-01 11 0.07 |
| 90 | (3, 2) | 2.02E-02 136 2.11 | 4.33E-01 14 0.11 |
| 100| (3, 2) | 2.57E-02 96 1.84 | 4.23E-01 10 0.09 |
| 80 | (3, 3) | 1.39E-02 35 0.34 | 1.41E+00 2 0.02 |
| 100| (3, 3) | 2.08E-02 89 1.69 | 1.41E+00 2 0.03 |
| 10 | (4, 1) | 3.86E-02 64 0.08 | 4.12E-01 341 0.21 |
| 20 | (4, 1) | 7.98E-02 40 0.16 | 4.45E-01 613 1.02 |
| 30 | (4, 1) | 7.37E-02 28 0.71 | 4.25E-01 485 6.55 |
| 40 | (4, 1) | 5.08E-02 25 1.62 | 4.47E-01 637 16.68 |
| 10 | (4, 2) | 4.98E-02 75 0.09 | 4.56E-01 299 0.19 |
| 20 | (4, 2) | 1.11E-01 53 0.20 | 4.73E-01 527 0.94 |
| 30 | (4, 2) | 7.33E-02 36 1.09 | 4.76E-01 394 6.06 |
| 40 | (4, 2) | 6.85E-02 27 1.75 | 4.70E-01 705 19.25 |
| 10 | (4, 3) | 9.57E-02 100 0.12 | 4.83E-01 664 0.41 |
| 20 | (4, 3) | 8.60E-02 69 0.27 | 5.00E-01 707 1.04 |
| 30 | (4, 3) | 1.29E-01 35 0.98 | 5.18E-01 645 9.72 |
| 40 | (4, 3) | 1.40E-01 30 1.86 | 5.41E-01 878 22.68 |
| 10 | (5, 2) | 5.36E-02 42 1.26 | 3.77E-01 13 0.04 |
| 20 | (5, 2) | 3.08E-02 34 4.11 | 4.72E-01 8 0.63 |
| 30 | (5, 2) | 3.05E-02 34 23.07 | 4.50E-01 7 3.42 |
| 10 | (5, 3) | 6.87E-02 41 1.17 | 3.63E-01 12 0.04 |
| 20 | (5, 3) | 2.08E-02 34 4.82 | 4.96E-01 10 0.85 |
| 30 | (5, 3) | 6.30E-02 35 20.89 | 5.15E-01 7 3.39 |
| 10 | (5, 4) | 5.00E-02 43 1.17 | 4.24E-01 16 0.04 |
| 20 | (5, 4) | 4.68E-02 34 4.47 | 5.05E-01 9 0.62 |
| 30 | (5, 4) | 8.16E-02 35 21.59 | 5.40E-01 7 4.16 |
Table 2 Comparison of HQ-ADMM for (2.4) and ALS for (2.2) when the ground truth tensor is contaminated by outliers

| n  | (d, t) | HQ-ADMM for (2.4) | ALS for (2.2) |
|----|--------|-------------------|---------------|
|    |        | err. | iter. | time | err. | iter. | time |
| 10 | (3, 1) | 4.54E-01 | 89 | 0.04 | 1.40E+00 | 150 | 0.04 |
| 20 | (3, 1) | 5.95E-02 | 46 | 0.04 | 1.41E+00 | 251 | 0.09 |
| 50 | (3, 1) | 1.99E-02 | 31 | 0.10 | 1.41E+00 | 757 | 0.95 |
| 80 | (3, 1) | 2.21E-02 | 27 | 0.55 | 1.41E+00 | 1456 | 12.17 |
| 90 | (3, 1) | 3.52E-02 | 28 | 0.70 | 1.41E+00 | 1204 | 11.59 |
| 100| (3, 1) | 2.82E-02 | 31 | 0.91 | 1.41E+00 | 1390 | 15.44 |
| 10 | (3, 2) | 4.32E-01 | 56 | 0.03 | 1.41E+00 | 120 | 0.04 |
| 20 | (3, 2) | 6.13E-01 | 35 | 0.04 | 1.41E+00 | 314 | 0.15 |
| 50 | (3, 2) | 7.50E-03 | 25 | 0.07 | 1.41E+00 | 592 | 0.69 |
| 80 | (3, 2) | 7.40E-03 | 25 | 0.42 | 1.41E+00 | 820 | 6.05 |
| 90 | (3, 2) | 6.66E-03 | 26 | 0.65 | 1.41E+00 | 828 | 7.80 |
| 100| (3, 2) | 8.16E-03 | 27 | 0.90 | 1.41E+00 | 928 | 11.99 |
| 10 | (3, 3) | 6.08E-03 | 25 | 0.42 | 1.41E+00 | 2 | 0.02 |
| 100| (3, 3) | 6.72E-03 | 27 | 0.80 | 1.41E+00 | 2 | 0.04 |
| 10 | (4, 1) | 1.04E-01 | 76 | 0.23 | 1.42E+00 | 187 | 0.14 |
| 20 | (4, 1) | 2.91E-02 | 34 | 0.28 | 1.41E+00 | 439 | 1.02 |
| 30 | (4, 1) | 4.40E-02 | 28 | 1.06 | 1.41E+00 | 1173 | 18.40 |
| 40 | (4, 1) | 6.09E-02 | 27 | 2.00 | 1.41E+00 | 885 | 26.09 |
| 10 | (4, 2) | 1.31E-01 | 67 | 0.08 | 1.41E+00 | 246 | 0.16 |
| 20 | (4, 2) | 5.23E-02 | 28 | 0.13 | 1.41E+00 | 729 | 1.12 |
| 30 | (4, 2) | 6.17E-02 | 27 | 0.85 | 1.41E+00 | 697 | 12.68 |
| 40 | (4, 2) | 3.36E-02 | 29 | 1.88 | 1.41E+00 | 1047 | 29.12 |
| 10 | (4, 3) | 1.40E-01 | 64 | 0.08 | 1.41E+00 | 208 | 0.13 |
| 20 | (4, 3) | 8.14E-02 | 29 | 0.12 | 1.41E+00 | 622 | 0.92 |
| 30 | (4, 3) | 8.45E-02 | 38 | 1.15 | 1.41E+00 | 900 | 14.85 |
| 40 | (4, 3) | 1.13E-01 | 30 | 2.12 | 1.41E+00 | 846 | 24.38 |
| 10 | (5, 2) | 1.05E-02 | 25 | 0.76 | 1.41E+00 | 336 | 0.71 |
| 20 | (5, 2) | 4.04E-02 | 30 | 3.82 | 1.41E+00 | 510 | 30.94 |
| 30 | (5, 2) | 9.72E-02 | 32 | 21.69 | 1.41E+00 | 755 | 294.56 |
| 10 | (5, 3) | 1.35E-02 | 25 | 0.14 | 1.41E+00 | 227 | 0.42 |
| 20 | (5, 3) | 3.03E-02 | 31 | 3.10 | 1.41E+00 | 511 | 29.23 |
| 30 | (5, 3) | 1.24E-01 | 29 | 19.92 | 1.41E+00 | 702 | 300.21 |
| 10 | (5, 4) | 1.35E-02 | 27 | 0.18 | 1.41E+00 | 349 | 0.74 |
| 20 | (5, 4) | 6.67E-02 | 29 | 2.87 | 1.41E+00 | 429 | 24.56 |
| 30 | (5, 4) | 1.07E-01 | 28 | 22.92 | 1.41E+00 | 623 | 289.96 |
ALS for (2.2) is not resistant with Cauchy noise as “err.” is large. Considering the efficiency, we see that HQ-ADMM all converges within 500 iterates and the CPU time is not large. Comparing with ALS, when \( d = 3 \) and 5, ALS is more efficient

### Table 3 Comparison of HQ-ADMM for (2.4) and ALS for (2.2) when the ground truth tensor is contaminated by Gaussian noise

| n  | (d, t) | HQ-ADMM for (2.4) | ALS for (2.2) |
|----|-------|------------------|--------------|
|    |       | err. iter. time  | err. iter. time |
| 10 | (3, 1) | 4.51E-02 198 0.09 | 4.09E-02 676 0.18 |
| 20 | (3, 1) | 3.62E-02 53 0.04 | 2.73E-02 564 0.19 |
| 50 | (3, 1) | 2.24E-02 30 0.08 | 2.18E-02 550 0.58 |
| 80 | (3, 1) | 2.14E-02 34 0.57 | 2.72E-02 716 5.78 |
| 90 | (3, 1) | 2.70E-02 33 0.79 | 2.44E-02 696 6.69 |
| 100| (3, 1) | 2.79E-02 34 0.98 | 2.28E-02 712 7.75 |
| 10 | (3, 2) | 3.89E-02 296 0.13 | 3.48E-02 16 0.01 |
| 20 | (3, 2) | 2.15E-02 65 0.05 | 1.87E-02 17 0.01 |
| 50 | (3, 2) | 7.99E-03 24 0.07 | 7.67E-03 14 0.02 |
| 80 | (3, 2) | 4.90E-03 24 0.40 | 4.82E-03 20 0.15 |
| 90 | (3, 2) | 4.68E-03 25 0.60 | 4.34E-03 41 0.40 |
| 100| (3, 2) | 3.85E-03 24 0.72 | 3.85E-03 7 0.10 |
| 10 | (4, 1) | 1.01E-01 673 0.83 | 8.62E-02 613 0.42 |
| 20 | (4, 1) | 7.46E-02 67 0.31 | 6.21E-02 699 1.33 |
| 30 | (4, 1) | 6.22E-02 29 1.05 | 6.61E-02 692 11.90 |
| 40 | (4, 1) | 8.68E-02 27 1.92 | 1.11E-01 858 24.49 |
| 10 | (4, 2) | 1.39E-02 45 0.15 | 1.74E-02 20 0.02 |
| 20 | (4, 2) | 4.75E-03 23 0.20 | 9.09E-03 17 0.05 |
| 30 | (4, 2) | 5.42E-03 26 0.91 | 2.71E-03 14 0.25 |
| 40 | (4, 2) | 2.26E-03 26 2.10 | 1.96E-03 41 1.24 |
| 10 | (4, 3) | 1.29E-02 48 0.17 | 1.23E-02 10 0.01 |
| 20 | (4, 3) | 4.93E-03 24 0.21 | 4.73E-03 10 0.04 |
| 30 | (4, 3) | 2.72E-03 25 0.98 | 2.88E-03 30 0.53 |
| 40 | (4, 3) | 1.95E-03 26 2.15 | 1.92E-03 21 0.67 |
| 10 | (5, 2) | 2.89E-02 35 0.15 | 2.60E-02 25 0.06 |
| 20 | (5, 2) | 7.81E-03 26 2.60 | 7.08E-03 39 2.42 |
| 30 | (5, 2) | 7.66E-03 26 21.18 | 4.14E-03 45 22.60 |
| 10 | (5, 3) | 1.64E-02 37 0.16 | 9.54E-03 12 0.04 |
| 20 | (5, 3) | 4.31E-03 25 2.52 | 3.69E-03 23 1.42 |
| 30 | (5, 3) | 2.14E-03 26 21.29 | 1.20E-03 21 10.02 |
| 10 | (5, 4) | 1.08E-02 36 0.16 | 6.41E-03 12 0.04 |
| 20 | (5, 4) | 7.04E-03 27 2.73 | 7.06E-03 16 1.06 |
| 30 | (5, 4) | 6.01E-03 25 20.30 | 4.35E-03 19 9.54 |
in most cases, while HQ-ADMM outperforms ALS when \( d = 4 \). This phenomenon still needs further research.

The cases contaminated by outliers are reported in Table 2, from which we can still observe that HQ-ADMM for solving (2.4) is consistent with outliers, owing to the redescending property of the Cauchy loss. HQ-ADMM outperforms ALS in terms of the iterates and CPU time.

The cases with Gaussian noise are reported in Table 3. It is known that model (2.2) is consistent with Gaussian noise, which can be seen from the table. We also observe that (2.4) is consistent with Gaussian noise from the third column, although the results are slightly worse than (2.2), as reported in the table. However, it is interesting to see that in some cases, namely, \((n, d, t) = (80, 3, 1), (30, 4, 1), (40, 4, 1), (20, 4, 2), (30, 4, 3), (20, 5, 4)\), HQ-ADMM for (2.4) is slightly better than ALS for (2.2). HQ-ADMM still shows its efficiency, and is more stable than ALS, as ALS needs much more iterates when \( t = 1 \).

**Simultaneous foreground-background extraction and compression**

Foreground-background extraction finds applications in video surveillance, where the aim is to detect moving objects such as human beings from static background. As the background changes little in the video, it is reasonable to project the background frames to a low dimensional subspace to compress the data. We show how this problem can be fitted into our model (2.4). Assume that a gray video consists of \( l \) frames, each of size \( m \times n \), resulting into a third-order tensor \( A \in \mathbb{R}^{l \times m \times n} \). Let \( A_i \) denote its \( i \)th frame. Our goal is to decompose it as \( A_r = B_r + F_r \), in which \( B_r \) and \( F_r \) denote the back-/foreground frames, respectively. Under the assumption that \( B_r \)'s lie in a low dimensional subspace with commonalities, we write \( B_r = U D_r V^T = \sum_{i=1}^{R} (D_r)_{ii} u_i v_i^T \), \( 1 \leq r \leq l \), where \( U = [u_1, \ldots, u_R] \), \( V = [v_1, \ldots, v_R] \) are orthonormal matrices, \( D_r \) is diagonal, and \( R \) is a parameter. On the other hand, the foreground is often sparse and can be recognized as outliers. Therefore, the Cauchy loss can be employed to control the effect of outliers. Denoting

\[
\phi_\delta(A_r - U D_r V^T) := \sum_{s=1}^{m} \sum_{t=1}^{n} \frac{\delta^2}{2} \log \left( 1 + \left( (A_r)_{st} - (U D_r V^T)_{st} \right)^2 / \delta^2 \right),
\]

**Table 4** HQ-ADMM for video surveillance with different \( R \)

| \( R \) | iter. | time | \( \frac{R(1000+144+176)}{1000+144+176} \) |
|--------|-------|------|----------------------------------|
| 10     | 43    | 33.86| 0.05%                            |
| 20     | 31    | 26.02| 0.1%                             |
| 30     | 26    | 21.58| 0.16%                            |
| 40     | 43    | 38.13| 0.21%                            |
| 50     | 31    | 28.78| 0.26%                            |

The last column shows the compressed ratio of the compressed background factors \( D, U, V \) to the sum of background frames \( B_r, 1 \leq r \leq l \).
the problem can be modeled as

$$\min_{U\geq I, V \geq I} U^t U = I, V^t V = I \sum_{r=1}^{l} \phi_{\delta}(A_r - UD_r V^t).$$

If we further denote $D \in \mathbb{R}^{l \times R}$ where the $r$th row is exactly the diagonal entries of $D_r$, the it can be written in the form of (2.4), i.e.,

$$\min_{U \geq I, V \geq I} \Phi_{\delta} \left( A - [D, U, V] \right), \tag{5.1}$$

where $\sigma$ is absorbed into $D$.

The tested video “airport” was downloaded from http://perception.i2r.a-star.edu.sg/bk/model/bk_index.html. The video consists of 4583 frames, each of size 144 $\times$ 176. We use 1000 frames, resulting into a tensor $A \in \mathbb{R}^{1000 \times 144 \times 176}$. $A$ is then normalized for conveniently choosing parameters, where we set $\delta = 0.05$, $\tau = 1$, and $\alpha = 10^{-8}$. The parameter $R$ varies in $\{10, 20, 30, 40, 50\}$. The quantitative results are reported in Table 4, in which we can see that HQ-ADMM stops around 30 $\sim$ 40 iterates, and consumes around 30 s, which demonstrates the efficiency of the algorithm. The last column shows the compressed ratio of the compressed background factors $D$, $U$, $V$ to the sum of background frames $B_r$, $1 \leq r \leq l$, from which we observe that the ratio is very small, resulting into low storage space. Some extracted frames with $R \in \{10, 30, 50\}$ are illustrated in Fig. 2. From the figures, we see that even when $R = 10$, HQ-ADMM can successfully separate the back-/foreground; of

![Fig. 2](image-url) Some extracting frames by HQ-ADMM from the video airport. Column (a): the original frames. Columns (b) and (c); extracted with $R = 10$. Columns (d) and (e): extracted with $R = 30$. Columns (f) and (g): extracted with $R = 50
course, when $R \geq 30$, the extracted frames are of higher quality, in that the background frames reconstructed from $UD_rV^\top$ are more clear. However, this also leads to the foreground might be recognized as the background, e.g., when $R = 50$, the person with a suitcase on top of the figures is partly recognized as the static background. This is because when the rank is higher, more information (foreground, or noise) is retained in the background $B_r$’s. Thus, how to choose a suitable $R$ still needs further research.

As a comparison, we also use the least squares–based problem (2.2) to model this problem and apply ALS to solve it, where we find that the least squares–based model does not perform well. To save space, we only illustrate the $R = 30$ case in Fig. 3. We can observe that two persons on top of the figures cannot be separated from the background. On the other hand, ALS cannot stop within 2000 iterates, with 723 s consumed. This shows that the derived model (5.1) and HQ-ADMM are more suitable for simultaneous foreground-background extraction and compression.

6 Conclusions

Heavy-tailed noise and outliers may contaminate real-world data. In the context of tensor canonical polyadic approximation problem with one or more latent factor matrices having orthonormal columns, most existing models are built upon the least squares loss, which is not resistant to heavy-tailed noise or outliers. To gain robustness, a Cauchy loss–based robust orthogonal tensor approximation model was proposed in this work. To efficiently solve this model, by exploring its half-quadratic property, HQ-ADMM was developed under the framework of alternating direction method of multipliers. Its global convergence was then established, thanks to some nice properties of the Cauchy loss. Numerical experiments on synthetic as well as real data demonstrate the effectiveness of the proposed model and algorithm. In future work, it would be interesting to incorporate other robust losses in the orthogonal tensor approximation problem and to apply HQ-ADMM to solve other Cauchy loss–based problems, as noted in Remark 3.2, and it would be important to investigate the convergence rate of HQ-ADMM.
Appendix

Proof of Theorem 4.1

To prove the convergence of a nonconvex ADMM, a key step is to upper bound the successive difference of the dual variables by the primal variables. Different from the nonconvex ADMMs in the literature, for HQ-ADMM, the weight $W^k$ brings barriers in the estimation of the upper bound. Fortunately, this can be overcome by realizing the relations between $W^k$, $T^k$ and $T^{k-1}$ based on Proposition 2.3, which will be given in Lemma A.1. With the upper bound at hand, we can derive the decreasing inequality with respect to $\{L^{k+1}T\}$ (Lemma A.2), whose verification is somewhat similar to that of a nonconvex block coordinate descent. Then, the boundedness of the variables is established in Theorem A.1. Key to the above two results is to set the parameter $\tau \geq \sqrt{10}$. Combining the above pieces, the subsequential convergence will be proved at the end of this subsection using a standard argument.

Lemma 2 It holds that

$$\|\Delta_{\gamma}^{k+1,k}\|_F \leq \|\Delta_{T}^{k+1,k}\|_F + \|\Delta_{T}^{k,k-1}\|_F.$$ 

Proof From (3.14), we have

$$W^k \otimes (T^{k+1} - A) + Y^k - \tau \left(\|\sigma^k; U^{k+1}\| - T^{k+1}\right) = 0,$$

which together with the definition of $Y^{k+1}$ yields

$$W^k \otimes (T^{k+1} - A) + Y^{k+1} = 0. \quad (33)$$

Therefore, we have

$$\begin{align*}
\|\Delta_{\gamma}^{k+1,k}\| &= \|W^k \otimes (T^{k+1} - A) - W^{k-1} \otimes (T^k - A)\|_F \\
&= \|W^k \otimes (T^{k+1} - A) - W^k \otimes (T^k - A) + W^k \otimes (T^k - A) - W^{k-1} \otimes (T^k - A)\|_F \\
&\leq \|W^k \otimes (T^{k+1} - T^k)\|_F + \|(W^k - W^{k-1}) \otimes (T^k - A)\|_F. \quad (34)
\end{align*}$$

Now, denote $E_1 := \|W^k \otimes (T^{k+1} - T^k)\|_F$ and $E_2 := \|(W^k - W^{k-1}) \otimes (T^k - A)\|_F$. We first consider $E_1$. From the definition of $W^k$, we easily see that $W^k_{i_1 \ldots i_d} \leq 1$ for each $i_1, \ldots, i_d$. Therefore,

$$E_1 \leq \|\Delta_{T}^{k+1,k}\|. \quad (35)$$

Next, we focus on $E_2$. To simplify notations we denote $a^k_{i_1 \ldots i_d} := T^k_{i_1 \ldots i_d} - A_{i_1 \ldots i_d}$ and

$$e_{i_1 \ldots i_d} := \delta^2 a^k_{i_1 \ldots i_d} \left(1 - \frac{1}{\delta^2 + \left(a^k_{i_1 \ldots i_d}\right)^2} - \frac{1}{\delta^2 + \left(a^{k-1}_{i_1 \ldots i_d}\right)^2}\right).$$
Then, $E_2$ can be expressed as

$$E_2^2 = \sum_{i_1=1,...,i_d=1}^{n_1,...,n_d} \left( \mathcal{W}^{k+1}_{i_1...i_d} - \mathcal{W}^k_{i_1...i_d} \right)^2 \left( \mathcal{T}^{k}_{i_1...i_d} - \mathcal{A}_{i_1...i_d} \right)^2$$

$$= \sum_{i_1=1,...,i_d=1}^{n_1,...,n_d} \left( \frac{1}{\delta^2 + (a^k_{i_1...i_d})^2} - \frac{1}{\delta^2 + (a^{k-1}_{i_1...i_d})^2} \right)^2 \delta^4 \left( a^k_{i_1...i_d} \right)^2$$

$$= \sum_{i_1=1,...,i_d=1}^{n_1,...,n_d} e^2_{i_1...i_d}.$$ 

It follows from Proposition 2.3 that

$$|e_{i_1...i_d}| \leq |a^k_{i_1...i_d} - a^{k-1}_{i_1...i_d}|,$$

and so

$$E_2 \leq \| \mathcal{T}^k - \mathcal{A} - (\mathcal{T}^{k-1} - \mathcal{A}) \|_F = \| \Delta_{\mathcal{T}}^{k-1} \|_F. \quad (36)$$

(.34) combining with (.35) and (.36) yields the desired result. \qed

With Lemma A.1, we then establish a decreasing inequality with respect to $\{\hat{L}_{\tau}^{k+1,k}\}$ defined in (4.1):

$$\hat{L}_{\tau}^{k+1,k} := L_\tau(\sigma^{k+1}, \mathcal{U}^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^{k+1}, \mathcal{T}^k), + \frac{2}{\tau} \| \mathcal{T} - \mathcal{T}' \|_F^2.$$ 

Key to the validness of the decreasing inequality is to set $\tau \geq \sqrt{10}$.

**Lemma 3** Let the parameter $\tau$ satisfy $\tau \geq \sqrt{10}$. Then, there holds

$$\hat{L}_\tau^{k,k-1} - \hat{L}_\tau^{k+1,k} \geq \alpha \sum_{j=1}^{d} \left\| \Delta_{U_j}^{k+1,k} \right\|_F^2 + \frac{1}{\tau} \left\| \Delta_{\mathcal{T}}^{k+1,k} \right\|_F^2, \forall k \geq 1,$$

where $\alpha > 0$ is defined in (3.11) and (3.12).
Proof We first consider the decrease caused by $U_j$. When $1 \leq j \leq d - t$, according to the algorithm, the expression of $L_\tau(\cdot)$, that $\|u_{j,i}^{k}\| = 1$ and recalling the definition of $u_{j,i}^{k+1}$, $v_{j,i}^{k+1}$, and $v_{j,i}^{k}$, we have

\[
L_\tau(\sigma^k, U_{j-1}^{k+1}, \ldots, U_j^{k+1}; U_j, \ldots, U_d^k; T^k, Y^k, W^k) - L_\tau(\sigma, U_{j-1}^{k+1}, \ldots, U_j^{k+1}; U_j, \ldots, U_d^k; T^k, Y^k, W^k)
\]

\[
= \frac{R}{\alpha} \sum_{i=1}^{R} \left( \sigma^k_i \cdot (Y^k + \tau T^k) u_{j,i}^{k+1} \otimes \cdots \otimes u_{j-1,i}^{k+1} \otimes u_{j,i}^k \otimes \cdots \otimes u_{d,i}^k, u_{j,i}^{k+1} - u_{j,i}^k \right)
\]

\[
= \sum_{i=1}^{R} \left( \sigma^k_i \cdot v_{j,i}^{k+1}, u_{j,i}^{k+1} - u_{j,i}^k \right)
\]

\[
= \sum_{i=1}^{R} \left( \sigma^k + \alpha u_{j,i}^k, u_{j,i}^{k+1} - u_{j,i}^k \right) + \frac{\alpha}{2} \| u_{j,i}^{k+1} - u_{j,i}^k \|^2
\]

\[
\geq \frac{\alpha}{2} \sum_{i=1}^{R} \| u_{j,i}^{k+1} - u_{j,i}^k \|^2 = \frac{\alpha}{2} \left\| \Delta U_j \right\|_F^2,
\]

where the fourth equality follows from the definition of $u_{j,i}^{k+1}$ and $v_{j,i}^{k+1}$, and the inequality is due to $\|v\| \geq (v, u)$ for any vectors $u$, $v$ of the same size with $\|u\| = 1$.

The decrease of $U_j$ when $d - t + 1 \leq j \leq d$ is similar. From the definition of $v_{j,i}^{k+1}$, it holds that

\[
L_\tau(\sigma^k, U_{j-1}^{k+1}, \ldots, U_j^{k+1}; U_j, \ldots, U_d^k; T^k, Y^k, W^k) - L_\tau(\sigma, U_{j-1}^{k+1}, \ldots, U_j^{k+1}; U_j, \ldots, U_d^k; T^k, Y^k, W^k)
\]

\[
= \sum_{i=1}^{R} \left( \sigma^k_i \cdot (Y^k + \tau T^k) u_{j,i}^{k+1} \otimes \cdots \otimes u_{j-1,i}^{k+1} \otimes u_{j,i}^k \otimes \cdots \otimes u_{d,i}^k, u_{j,i}^{k+1} - u_{j,i}^k \right)
\]

\[
= \left( V_j^{k+1} \cdot \text{diag}(\sigma^k), U_j^{k+1} - U_j^k \right)
\]

\[
= \left( V_j^{k+1} \cdot \text{diag}(\sigma^k) + \alpha U_j^k, U_j^{k+1} - U_j^k \right) + \frac{\alpha}{2} \left\| U_j^{k+1} - U_j^k \right\|_F^2
\]

\[
\geq \frac{\alpha}{2} \left\| \Delta U_j \right\|_F^2.
\]
where the inequality follows from the definition of $U^{k+1}_j$ in (3.12). To show the decrease of $\mathcal{T}$, note that $L_T(\cdot)$ is strongly convex with respect to $\mathcal{T}$, based on which we can easily deduce that

$$L_T(\sigma^k, U^{k+1}, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k) - L_T(\sigma^k, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^k, \mathcal{W}^k) \geq \frac{\tau}{2} \left\| \Delta_{\mathcal{T}}^{k+1,k} \right\|^2_F. \tag{.39}$$

Next, it follows from the definition of $\mathcal{Y}^{k+1}$ and Lemma A.1 that

$$L_T(\sigma^k, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^k, \mathcal{W}^k) - L_T(\sigma^k, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^k) = \left\langle \Delta_{\mathcal{Y}}^{k+1,k}, \left\| \sigma^k; U^{k+1} \right\| - \mathcal{T}^{k+1} \right\rangle = - \frac{1}{\tau} \left\| \Delta_{\mathcal{Y}}^{k+1,k} \right\|^2_F \geq - \frac{2}{\tau} \left( \left\| \Delta_{\mathcal{T}}^{k+1,k} \right\|^2_F + \left\| \Delta_{\mathcal{T}}^{k,k-1} \right\|^2_F \right). \tag{.40}$$

Finally, it follows from the definition of $\sigma^{k+1}$ and $\mathcal{W}^{k+1}$ that

$$L_T(\sigma^k, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^k) - L_T(\sigma^{k+1}, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^k) \geq 0, \tag{.41}$$

$$L_T(\sigma^{k+1}, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^k) - L_T(\sigma^{k+1}, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^{k+1}) \geq 0. \tag{.42}$$

As a result, summing up (.37)–(.42) yields

$$L_T(\sigma^k, U^k, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k) - L_T(\sigma^{k+1}, U^{k+1}, \mathcal{T}^{k+1}, \mathcal{Y}^{k+1}, \mathcal{W}^{k+1}) \geq \frac{\alpha}{2} \sum_{j=1}^d \left\| \Delta_{\mathcal{U}_j}^{k+1,k} \right\|^2_F + \left( \frac{\tau}{2} - \frac{2}{\tau} \right) \left\| \Delta_{\mathcal{T}}^{k+1,k} \right\|^2_F - \frac{2}{\tau} \left\| \Delta_{\mathcal{T}}^{k,k-1} \right\|^2_F \geq \frac{\alpha}{2} \sum_{j=1}^d \left\| \Delta_{\mathcal{U}_j}^{k+1,k} \right\|^2_F + \left( \frac{2}{\tau} + \frac{1}{\tau} \right) \left\| \Delta_{\mathcal{T}}^{k+1,k} \right\|^2_F - \frac{2}{\tau} \left\| \Delta_{\mathcal{T}}^{k,k-1} \right\|^2_F. \tag{.43}$$

where the last inequality follows from the range of $\tau$. Rearranging the terms of (.43) gives the desired results. This completes the proof. \hfill \square

We then show that $\tilde{L}^{k,k-1}_t$ defined in Lemma A.2 is lower bounded and the sequence $\{\sigma^k, U^k, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k\}$ is bounded as well.

**Theorem 3** Under the setting of Lemma A.2, $\{\tilde{L}^{k,k-1}_t\}$ is bounded. The sequence $\{\sigma^k, U^k, \mathcal{T}^k, \mathcal{Y}^k, \mathcal{W}^k\}$ generated by Algorithm 1 is bounded as well.
Proof Denote $Q^k(\cdot) := \frac{1}{2} \left\| \sqrt{\mathcal{W}^k} \otimes (\cdot - A) \right\|_F^2$; thus, we have $\nabla Q^k(\mathcal{T}) = \mathcal{W}^k \otimes (\mathcal{T} - A)$, and it then follows from the quadraticity of $Q^k(\cdot)$ and $\mathcal{V}^k = -\mathcal{W}^{k-1} \otimes (\mathcal{T}^k - A)$ from (43) that

\[
Q^{k-1}(\mathcal{T}^k) - Q^{k-1}(\mathcal{V}^k) = \langle \mathcal{W}^{k-1} \otimes (\mathcal{V}^k - A), \mathcal{T}^k - \mathcal{V}^k \rangle \]

\[
+ \frac{1}{2} \left\| \sqrt{\mathcal{W}^{k-1}} \otimes (\mathcal{V}^k - \mathcal{T}^k) \right\|_F^2 - \left\langle \mathcal{V}^k, \left[ \mathcal{W}^k \right] - \mathcal{T}^k \right\rangle
\]

\[
= \frac{1}{2} \left\| \sqrt{\mathcal{W}^{k-1}} \otimes (\mathcal{V}^k - \mathcal{T}^k) \right\|_F^2

+ \left\langle \mathcal{W}^{k-1} \otimes (\mathcal{V}^k - A) - \mathcal{W}^{k-1} \otimes (\mathcal{T}^k - A), \mathcal{T}^k - \mathcal{V}^k \right\rangle
\]

\[
= -\frac{1}{2} \left\| \mathcal{W}^{k-1} \otimes (\mathcal{V}^k - \mathcal{T}^k) \right\|_F^2
\]

where the last inequality uses the fact that $0 < \mathcal{W}^{k-1}_{i_1 \cdots i_d} \leq 1$. It thus follows that for any $k \geq 2$,

\[
\tilde{L}^{k-1,k-2} = \tilde{L}_\tau(\mathcal{V}^{k-1}, \mathcal{U}^{k-1}, \mathcal{T}^{k-1}, \mathcal{V}^{k-1}, \mathcal{W}^{k-1}, \mathcal{T}^{k-1})
\]

\[
\geq \tilde{L}_\tau(\mathcal{V}^k, \mathcal{U}^k, \mathcal{T}^k, \mathcal{V}^k, \mathcal{W}^{k-1}, \mathcal{T}^{k-1})
\]

\[
= Q^{k-1}(\mathcal{T}^k) + \frac{\delta^2}{2} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \phi(\mathcal{W}^{k-1}_{i_1 \cdots i_d}) - \left\langle \mathcal{V}^k, [\mathcal{W}^k] - \mathcal{T}^k \right\rangle
\]

\[
+ \frac{\tau}{2} \left\| [\mathcal{W}^k] - \mathcal{T}^k \right\|_F^2 + \frac{2}{\tau} \left\| \Delta^{k-1}_k \right\|_F^2
\]

\[
\geq Q^{k-1}(\mathcal{V}^k) + \frac{\tau - 1}{2} \left\| [\mathcal{W}^k] - \mathcal{T}^k \right\|_F^2
\]

\[
+ \frac{\delta^2}{2} \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \phi(\mathcal{W}^{k-1}_{i_1 \cdots i_d}) + \frac{2}{\tau} \left\| \delta P k k - 1 \right\|_F^2
\]

\[
> -\infty,
\]

where the first inequality follows from the proof of Lemma A.2 (summing up (37)–(41), the second one comes from (44), and the last one is due to the range of $\tau$ and $\phi(\cdot) \geq 0$. Thus, $\{\tilde{L}_\tau^{k-1,k-1}\}$ is a lower bounded sequence. This together with Lemma A.2 shows that $\{\tilde{L}_\tau^{k,k-1}\}$ is bounded. We then show the boundedness of $\{\mathcal{V}^k, \mathcal{U}^k, \mathcal{T}^k, \mathcal{V}^k, \mathcal{W}^k\}$. The boundedness of $\{\mathcal{U}^k\}$ and $\{\mathcal{W}^k\}$ is obvious. Next, denote $g(\mathcal{V}^k)$ as the formulation in lines 5–6 of (45) with respect to $\mathcal{V}^k$. Proposition 2.1 shows that $\mathcal{W}^k \otimes \mathcal{U}^k$ is orthonormal and hence $\left\| [\mathcal{W}^k \otimes \mathcal{U}^k] - \mathcal{T}^k \right\|_F^2$ is strongly convex with respect to $\mathcal{V}^k$; this together with the convexity of $Q^{k-1}(\mathcal{V}^k)$ shows that $g(\mathcal{V}^k)$ is strongly convex with respect to $\mathcal{V}^k$. Combining this with (45) gives the boundedness of $\{\mathcal{V}^k\}$. Quite similarly, we have that $\{\mathcal{T}^k\}$ is bounded. Finally, the boundedness of $\{\mathcal{V}^k\}$ follows from the expression of the $\mathcal{T}$-subproblem (3.14). As a result, $\{\mathcal{V}^k, \mathcal{U}^k, \mathcal{T}^k, \mathcal{V}^k, \mathcal{W}^k\}$ is a bounded sequence. This completes the proof. \qed
Proof of Theorem 4.1  Lemma A.2 in connection with Theorem A.1 yields points 1, 2, and (4.2); (4.2) together with Lemma A.1 and the definition of \(\gamma^{k+1}, \sigma^{k+1}\), and \(\mathcal{W}^{k+1}\) gives (4.3). On the other hand, since the sequence is bounded, limit points exist. Assume that \(\{\sigma^*, U^*, T^*, Y^*, \mathcal{W}\}\) is a limit point with

\[
\lim_{l \to \infty} \{\sigma^{k_l}, U^{k_l}, T^{k_l}, Y^{k_l}, \mathcal{W}^{k_l}\} = \{\sigma^*, U^*, T^*, Y^*, \mathcal{W}\}.
\]

Therefore, taking the limit into \(l\) with respect to the \(u_{j,i}\)-subproblem (3.11) yields

\[
v_{j,i}^* \sigma_{i}^* + \alpha u_{j,i}^* = \|v_{j,i}^*\| u_{j,i}^*, \quad 1 \leq j \leq d-t, \quad 1 \leq i \leq R. \tag{46}
\]

Multiplying both sides by \(u_{j,i}^*\) gives

\[
\|v_{j,i}^*\| = \alpha + \sigma_{i}^* \langle v_{j,i}^*, u_{j,i}^* \rangle = \alpha + \sigma_{i}^* \left( \langle Y^* + \tau T^*, \bigotimes_{j=1}^{d} u_{j,i}^* \rangle \right) = \alpha + \tau (\sigma_{i}^*)^2, \tag{47}
\]

where the second equality follows from the definition of \(v_{j,i}^*\) and the last one is given by passing the limit into the expression of \(\sigma_{i}^{k_l+1}\) (3.15). Thus, (4.6) together with (4.7) gives

\[
\left( Y^* + \tau T^* \right) \bigotimes_{l \neq j}^{d} u_{l,i}^* = \sigma_{i}^* \tau u_{j,i}^*, \tag{48}
\]

i.e., the first equation of the stationary point system (3.10).

Taking the limit into \(l\) with respect to the \(U_{j}\)-subproblem (3.12) and noticing the expression (3.13), we get

\[
V_{j}^* \text{diag}(\sigma^*) + \alpha U_{j}^* = U_{j}^* H_{j}^*,
\]

where \(H_{j}^*\) is a symmetric matrix. Writing it columnwisely, we obtain

\[
\sigma_{i}^* \left( Y^* + \tau T^* \right) \bigotimes_{l \neq j}^{d} u_{l,i}^* = \sum_{i=1}^{R} (H_{j}^*)_{i,r} u_{j,r}^* - \alpha u_{j,r}^*, \quad d-t+1 \leq j \leq d, \quad 1 \leq i \leq R.
\]

Denoting \(\Lambda_{j}^* := H_{j}^* - \alpha I\), the above is exactly the third equality of (3.10). On the other hand, passing the limit into the expression of \(\mathcal{T}^*\) (3.14) and \(\mathcal{W}^*\) (3.16) respectively gives the \(T^*\) - and \(\mathcal{W}^*\)- formulas in (3.10). Finally, the first expression of (4.3) yields \(\mathcal{T}^* = \{\sigma^*; \mathcal{U}\}\). Taking the above pieces together, we have that \(\{\sigma^*, U^*, T^*, Y^*, \mathcal{W}\}\) satisfies the stationary point system (3.10).

Next, we show that \(\{\sigma^*, U^*\}\) is also a stationary point of problem (2.4). We define its Lagrangian function as \(L_{\Phi} := \Phi_{3}(\sigma, U) - \sum_{j,i=1}^{d-t,R} \eta_{j,i} \langle u_{j,i}, u_{j,i} - 1 \rangle - \sum_{j=d-t+1}^{d} \{\Lambda_{j}, U_{j}^{+} U_{j} - I\}\), similar to that in (3.5). Taking derivative yields

\[
\begin{aligned}
\partial_{u_{j,i}} \Phi_{3}(\sigma, U) &= \eta_{j,i} u_{j,i} \leftrightarrow \mathcal{W} \oplus (\{\sigma; U\} - \mathcal{A}) \cdot \sigma_{i} \bigotimes_{l \neq j}^{d} u_{j,i} = \eta_{j,i} u_{j,i}, \quad 1 \leq j \leq d-t, \quad 1 \leq i \leq R, \\
\partial_{u_{j,i}} \Phi_{3}(\sigma, U) &= \sum_{r=1}^{R} (\Lambda_{j})_{t,r} u_{j,r} \leftrightarrow \mathcal{W} \oplus (\{\sigma; U\} - \mathcal{A}) \cdot \sigma_{i} \bigotimes_{l \neq j}^{d} u_{j,r} = \sum_{r=1}^{R} (\Lambda_{j})_{t,r} u_{j,r}, \quad d-t+1 \leq j \leq d, \quad 1 \leq i \leq R, \\
\partial_{\sigma} \Phi_{3}(\sigma, U) &= 0 \leftrightarrow \left\{ \mathcal{W} \oplus (\{\sigma; U\} - \mathcal{A}) \cdot \bigotimes_{l=1}^{d} u_{l,i} \right\} = 0, \quad 1 \leq i \leq R.
\end{aligned}
\]

(49) Springer
where \( \mathcal{W}_{i_1 \ldots i_d} = \delta^2 \left( 1 + \left[ \sigma; U \right]_{i_1 \ldots i_d} - A_{i_1 \ldots i_d} \right)^2 / \delta^2 \); multiplying \( u_{j,i} \) in both sides of the first equality above, and noticing the last equality, we get \( \eta_{j,i} = 0 \). Since \( T^* = [\sigma^*; U^*] \), the \( T \)-subproblem (3.14) also gives \( \mathcal{V}^* = \mathcal{W}^* \otimes (\sigma^*; U^*) - A \). This together with (4.8) and that \( T^* \otimes_{i\neq j} u^*_{j,i} = [\sigma^*; U^*] \otimes_{i\neq j} u^*_{j,i} = \sigma^* u^*_{j,i} \) gives \( \mathcal{W}^* \otimes (\sigma^*, U^*) - A \otimes_{i\neq j} u^*_{j,i} = 0 \), i.e., the first equality of (4.9) by noticing \( \eta_{j,i} = 0 \). In a similar vein, we get that
\[
\sigma_i^* \mathcal{W}^* \otimes (\sigma^*; U^*) - A = \sum_{i=1}^R (H^*_j)_{i,r} u^*_{j,r} - (\alpha + \tau \sigma_i^*) u^*_{j,i}.
\]
Taking \( \Lambda_j := H^*_j - (\alpha + \tau \sigma_i^*) I \) gives the second relation of (4.9). The last equality follows directly from \( \mathcal{W}^* \otimes (\sigma^*, U^*) - A \otimes_{i\neq j} u^*_{j,i} = 0 \). The proof has been completed.

**Proof of Theorem 4.2**

To prove Theorem 4.2, we first recall some definitions from nonsmooth analysis. Denote \( \text{dom} f := \{ x \in \mathbb{R}^n \mid f(x) < +\infty \} \).

**Definition 1** (c.f. [2]) For \( x \in \text{dom} f \), the Fréchet subdifferential, denoted as \( \hat{\partial} f(x) \), is the set of vectors \( z \in \mathbb{R}^n \) satisfying
\[
\liminf_{\substack{y \to x \\ y \neq x}} \frac{f(y) - f(x) - \langle z, y-x \rangle}{\|y-x\|} \geq 0.
\]
The subdifferential of \( f \) at \( x \in \text{dom} f \), written \( \partial f \), is defined as
\[
\partial f(x) := \left\{ z \in \mathbb{R}^n : \exists x^k \to x, f \left( x^k \right) \to f(x), z^k \in \hat{\partial} f \left( x^k \right) \right\}.
\]

It is known that \( \hat{\partial} f(x) \subset \partial f(x) \) for each \( x \in \mathbb{R}^n \) [4]. An extended-real-valued function is a function \( f : \mathbb{R}^n \to [-\infty, \infty] \), which is proper if \( f(x) > -\infty \) for all \( x \), and \( f(x) < \infty \) for at least one \( x \). It is called closed if it is lower semi-continuous (l.s.c. for short). The global convergence relies on the the Kurdyka-Łojasiewicz (KL) property given as follows:

**Definition 2** (KL property and KL function, c.f. [2, 4]) A proper function \( f \) is said to have the KL property at \( \bar{x} \in \text{dom} \partial \hat{f} := \{ x \in \mathbb{R}^n \mid \partial \hat{f}(x) \neq \emptyset \} \), if there exist \( \bar{\varepsilon} \in (0, \infty] \), a neighborhood \( N \) of \( \bar{x} \), and a continuous and concave function \( \psi : (0, \bar{\varepsilon}) \to \mathbb{R}_+ \) which is continuously differentiable on \( (0, \bar{\varepsilon}) \) with positive derivatives and \( \psi(0) = 0 \), such that for all \( x \in N \) satisfying \( f(\bar{x}) < f(x) < f(\bar{x}) + \bar{\varepsilon} \), it holds that
\[
\psi'(f(x) - f(\bar{x})) \text{dist}(0, \partial \hat{f}(x)) \geq 1,
\]
where \( \text{dist}(0, \partial \hat{f}(x)) \) means the distance from the original point to the set \( \partial \hat{f}(x) \). If a proper and l.s.c. function \( f \) satisfies the KL property at each point of \( \text{dom} \partial \hat{f} \), then \( f \) is called a KL function.
We then simplify \( \tilde{L}_\tau(\cdot) \) by eliminating the variables \( W \) and \( \sigma \). First, from the definition of \( W^{k+1} \) and Lemma 2.1, we have that

\[
\| \sqrt{W}^{k+1} \otimes \left( \mathcal{T}^{k+1} - A \right) \|_F^2 + \delta^2 \sum_{i_1=1,\ldots,i_d=1}^{n_1,\ldots,n_d} \varphi(W_{i_1,\ldots,i_d}^{k+1}) = \Phi_\delta(T^{k+1} - A),
\]

where \( \Phi_\delta(\cdot) \) is defined in (2.4). This eliminate the \( W \) from \( \tilde{L}_\tau(\cdot) \). On the other hand, it follows from the definition of \( \sigma^{k+1} \) (3.15) that

\[
-\left( \gamma^{k+1}, [\sigma^{k+1}; U^{k+1}] - T^{k+1} \right) + \frac{\tau}{2} \left\| [\sigma^{k+1}; U^{k+1}] - T^{k+1} \right\|_F^2
\]

\[
= \left( \gamma^{k+1}, T^{k+1} \right) + \frac{\tau}{2} \left\| T^{k+1} \right\|_F^2 - \frac{1}{2\tau} \sum_{i=1}^R \left( \gamma^{k+1} + \tau T^{k+1} \right) \bigotimes_{j=1}^d u_{j,i}^{k+1}^2.
\]

Thus, \( \sigma \) is also eliminated. In what follows, whenever necessary, \( \sigma^i_j \) still represents the expression \( (\gamma^i + \tau T^i) \bigotimes_{j=1}^d u_{j,i}^i / \tau \), but we only treat it as a representation instead of a variable.

Then, \( \tilde{L}_\tau(\sigma^{k+1}, U^{k+1}, T^{k+1}, \gamma^{k+1}, W^{k+1}, T^{k}) \) can be equivalently written as

\[
\tilde{L}_\tau(U^{k+1}, T^{k+1}, \gamma^{k+1}, T)^k = \frac{1}{2} \Phi_\delta(T^{k+1} - A) + \left( \gamma^{k+1}, T^{k+1} \right) + \frac{\tau}{2} \left\| T^{k+1} \right\|_F^2
\]

\[
- \frac{1}{2\tau} \sum_{i=1}^R \left( \gamma^{k+1} + \tau T^{k+1} \right) \bigotimes_{j=1}^d u_{j,i}^{k+1}^2 + \frac{2}{\tau} \left\| A^{k+1,k} \right\|_F^2.
\]

In addition, we denote

\[
\tilde{L}_{\tau,\alpha}(U, T, \gamma, T') := \tilde{L}_\tau(U, T, \gamma, T') - \frac{\alpha}{2} \sum_{j=1}^d \| U_j \|^2 + \sum_{j=1}^{d-t} \sum_{i=1}^R \lambda_{s(n_j, i)}(u_{j,i}) + \sum_{j=d-t+1}^d \sum_{i=1}^R \lambda_{s(n_j, i)}(U_j).
\]

We can see that under the constraints of the optimization problem (2.4), \( \tilde{L}_{\tau,\alpha}(\cdot) = \tilde{L}_\tau(\cdot) - \frac{\alpha d R}{2} \). This together with Theorem 4.1 tells us that the sequence \( \{ \tilde{L}_{\tau,\alpha}(U^{k+1}, T^{k+1}, \gamma^{k+1}, T^{k}) \} \) is also bounded and nonincreasing. In addition, we have that \( \tilde{L}_{\tau,\alpha}(\cdot) \) is a KL function.

**Proposition .4** \( \tilde{L}_{\tau,\alpha}(U, T, \gamma, T') \) defined above is a proper, l.s.c., and KL function.

**Proof** It is clear that \( \tilde{L}_{\tau,\alpha}(\cdot) \) is proper and l.s.c.. Next, since the constrained sets in (2.4) are all Stiefel manifolds, items 2 and 6 of [4, Example 2] tell us that they are semi-algebraic sets, and their indicator functions are semi-algebraic functions. Therefore, the indicator functions are KL functions [4, Theorem 3]. On the other hand, the remaining part of \( \tilde{L}_{\tau,\alpha} \) (besides the indicator functions) is an analytic function and hence it is KL [4]. As a result, \( \tilde{L}_{\tau,\alpha}(U, T, \gamma, T') \) is a KL function. \( \square \)
In the sequel, we mainly rely on \( \hat{L}_{\tau, \alpha}(\cdot) \) to prove the global convergence. For convenience, we denote 
\[
\hat{L}_{\tau, \alpha}^{k+1, k} := \hat{L}_{\tau, \alpha}(U^{k+1}, T^{k+1}, Y^{k+1}, T^k), \quad \text{and} \\
\delta \hat{L}_{\tau, \alpha}^{k+1, k} := \partial \hat{L}_{\tau, \alpha}(U^{k+1}, T^{k+1}, Y^{k+1}, T^k);
\]
denote \( \Delta U^{k+1, k} := (U^{k+1}, T^{k+1}) - (U^k, T^k) \), and
\[
\| \Delta U^{k+1, k} \|_F := \sqrt{\sum_{j=1}^d \| \Delta U_j^{k+1, k} \|_F^2 + \| \Delta T^k \|_F^2}.
\]

**Lemma 4** There exists a large enough constant \( c_0 > 0 \), such that
\[
dist(0, \delta \hat{L}_{\tau, \alpha}^{k+1, k}) \leq c_0 \left( \| \Delta U^{k+1, k} \|_F + \| \Delta T^k \|_F \right). \tag{A.19}
\]

**Proof** We first consider \( \partial_u \hat{L}_{\tau, \alpha}^{k+1, k}, 1 \leq j \leq d - t, 1 \leq i \leq R, \) and \( \partial_u \hat{L}_{\tau, \alpha}^{k+1, k}, d - t + 1 \leq j \leq d, \) respectively. In what follows, we denote
\[
\bar{v}^{k+1}_{j,i} := \sigma_i^{k+1} \left( Y^{k+1} + \tau T^{k+1} \right) \bigotimes_{l \neq j} u_{l,i}^{k+1} + \alpha u_{j,i}^{k+1}, \quad \text{and} \\
\bar{v}_{j,i}^{k+1} := [\bar{v}^{k+1}_{j,1}, \ldots, \bar{v}^{k+1}_{j,R}]\quad \text{for later use. In addition, denote} \\
\bar{v}^{k+1}_j := [\bar{v}^{k+1}_{j,1}, \ldots, \bar{v}^{k+1}_{j,R}].
\]

For \( 1 \leq j \leq d - t, \) one has
\[
\partial_{u_{j,i}} \hat{L}_{\tau, \alpha}^{k+1, k} = -\sigma_i^{k+1} \left( Y^{k+1} + \tau T^{k+1} \right) \bigotimes_{l \neq j} u_{l,i}^{k+1} - \alpha u_{j,i}^{k+1} + \partial_{\text{st} n_{j,i}} (u_{j,i}^{k+1})
\]
\[
= -\bar{v}_{j,i}^{k+1} + \partial_{\text{st} n_{j,i}} \bar{v}_{j,i}^{k+1}. \tag{A.20}
\]

We then wish to show that
\[
\bar{v}_{j,i}^{k+1} \in \hat{\partial}_{\text{st} n_{j,i}} (u_{j,i}^{k+1}) \subset \partial_{\text{st} n_{j,i}} (u_{j,i}^{k+1}). \tag{A.21}
\]

The proof is similar to that of [53, Lemma 6.1]. First, from the definition of \( \text{st} n_{j,i}(\cdot) \) and \( \hat{\partial}_{\text{st} n_{j,i}} (\cdot) \) in (.50), it is not hard to see that if \( y \notin \text{st} n_{j,i} \), then (.50) clearly holds when \( z = \bar{v}_{j,i}^{k+1} \); otherwise if \( y \in \text{st} n_{j,i} \), i.e., \( \| y \| = 1 \), then from the definition of \( u_{j,i}^{k+1} \), we see that
\[
u_{j,i}^{k+1} = \arg \max_{\| y \| = 1} \langle y, \bar{v}_{j,i}^{k+1} \rangle \Leftrightarrow \langle \bar{v}_{j,i}^{k+1}, u_{j,i}^{k+1} - y \rangle \geq 0, \quad \forall \| y \| = 1,
\]
which together with \( \text{st} n_{j,i}(y) = 0 \) and \( \text{st} n_{j,i}(u_{j,i}^{k+1}) = 0 \) gives
\[
\liminf_{y \neq u_{j,i}^{k+1}, y \to u_{j,i}^{k+1}} \frac{\text{st} n_{j,i}(y) - \text{st} n_{j,i}(u_{j,i}^{k+1}) - \langle \bar{v}_{j,i}^{k+1}, y - u_{j,i}^{k+1} \rangle}{\| y - u_{j,i}^{k+1} \|} \geq 0.
\]
As a result, (A.21) is true, which together with (A.20) shows that
\[
\bar{v}_{j,i}^{k+1} - \bar{v}_{j,i}^{k+1} \in \partial_{u_{j,i}} \hat{L}_{\tau, \alpha}^{k+1, k}, \quad 1 \leq j \leq d - t, \quad 1 \leq i \leq R.
\]

Let \( 0 \) denote the origin. Then by using the triangle inequality and the boundedness of \( \{a^k, U^k, T^k, Y^k\} \), and noticing the definition of \( \Delta U^{k+1, k} \), there must exist
large enough constants $c_1, c_2 > 0$ only depending on $\tau, \alpha$, and the size of \( \{\sigma^k, U^k, T^k, Y^k\} \), such that

\[
\text{dist}(0, \partial_{u_{j,i}} \tilde{L}_{r,\alpha}^{k+1,k}) \\
\leq \left\| \tilde{V}_{k-1}^{j,i} - \tilde{V}_{k+1}^{j,i} \right\| \\
\leq c_1 \left( \sum_{j=1}^{d} \left\| \Delta U_{k, j}^{k+1,k} \right\|_F + \left\| \Delta T_{k}^{k+1,k} \right\|_F + \left\| \Delta Y_{k}^{k+1,k} \right\|_F \right) \\
\leq c_1 \left( \sum_{j=1}^{d} \left\| \Delta U_{k, j}^{k+1,k} \right\|_F + 2 \left\| \Delta T_{k}^{k+1,k} \right\|_F + \left\| \Delta Y_{k}^{k+1,k} \right\|_F \right) \\
\leq c_2 \left( \left\| \Delta U_{T, k}^{k+1,k} \right\|_F + \left\| \Delta Y_{T, k}^{k+1,k} \right\|_F \right), \quad 1 \leq j \leq d - t. \quad (A.22)
\]

On the other hand, for $d - t + 1 \leq j \leq d$, by noticing the definition of $\tilde{V}_{k+1}^{j}$, we have

\[
\partial_{u_{j,i}} \tilde{L}_{r,\alpha}^{k+1,k} = -\tilde{V}_{k+1}^{j} + \partial_{\text{st}}(n_j, R)(U_{k+1}^{j}).
\]

From the definition of $U_{k+1}^{j}$ in (3.12) and similar to the above argument, we can show that $\tilde{V}_{k+1}^{j} \in \partial_{\text{st}}(n_j, R)(U_{k+1}^{j})$. Thus,

\[
\tilde{V}_{k+1}^{j} - \tilde{V}_{j}^{k+1} \in \partial_{u_{j,i}} \tilde{L}_{r,\alpha}^{k+1,k}, \quad d - t + 1 \leq j \leq d.
\]

Similar to (A.22), there exists a large enough constant $c_3 > 0$ such that

\[
\text{dist}(0, \partial_{u_{j,i}} \tilde{L}_{r,\alpha}^{k+1,k}) \leq c_3 \left( \left\| \Delta U_{T, k}^{k+1,k} \right\|_F + \left\| \Delta Y_{T, k}^{k+1,k} \right\|_F \right), \quad d - t + 1 \leq j \leq d. \quad (A.23)
\]

We then consider

\[
\nabla_{T} \tilde{L}_{r,\alpha}^{k+1,k} = \nabla_{\mathcal{K}}^{k+1} \otimes \left( T_{k+1}^{k+1} - A \right) + \nabla_{\mathcal{K}}^{k+1} - \Delta_{T}^{k+1} + \frac{\tau}{\tau} (T_{k+1}^{k+1} - T_{k}^{k+1}).
\]

Note that $\nabla_{\mathcal{K}}^{k+1}$ and $\Delta_{\mathcal{T}}^{k+1}$ above are only representations instead of variables, which represent (3.16) and (3.15). From the expression of $\nabla_{\mathcal{K}}^{k+1}$ in (3.33), we have

\[
\left\| \nabla_{\mathcal{K}}^{k+1} \otimes \left( T_{k+1}^{k+1} - A \right) + \nabla_{\mathcal{K}}^{k+1} \right\|_F = \left\| \left( \nabla_{\mathcal{K}}^{k+1} - \nabla_{\mathcal{K}}^{k} \right) \otimes \left( T_{k+1}^{k+1} - A \right) \right\|_F \\
\leq \left\| \Delta_{T}^{k+1} \right\|_F,
\]

where the inequality follows from Proposition 2.3. On the other side,

\[
\tau \left\| \left[ \sigma^{k+1}; U^{k+1} \right] - T_{k+1}^{k+1} \right\|_F = \tau \left\| \left[ \sigma^{k+1}; U^{k+1} \right] - \left[ \sigma^{k}; U^{k+1} \right] + \left[ \sigma^{k}; U^{k+1} \right] - T_{k+1}^{k+1} \right\|_F \\
\leq \tau \left\| \left[ \sigma^{k+1}; U^{k+1} \right] - \left[ \sigma^{k}; U^{k+1} \right] \right\|_F + \left\| \Delta_{T}^{k+1} \right\|_F \\
\leq c_4 \left( \left\| \Delta_{T}^{k+1} \right\|_F + \left\| \Delta_{T}^{k+1} \right\|_F \right), \quad (A.24)
\]

where \( c_4 > 0 \) is large enough. Combining the above pieces shows that there exists a large enough constant \( c_5 > 0 \) such that
\[
\left\| \nabla_T \tilde{L}^{k+1,k}_{\tau,\alpha} \right\|_F \leq c_5 \left( \left\| \Delta^{k+1,k}_{U,T} \right\|_F + \left\| \Delta^{k,k-1}_{U,T} \right\|_F \right).
\] (A.25)

Next, it follows from (A.24) that
\[
\left\| \nabla_Y \tilde{L}^{k+1,k}_{\tau,\alpha} \right\|_F = \left\| \sigma^{k+1} ; U^{k+1} - T^{k+1} \right\|_F \leq \frac{c_4}{\tau} \left( \left\| \Delta^{k+1,k}_{U,T} \right\|_F + \left\| \Delta^{k,k-1}_{U,T} \right\|_F \right).
\] (A.26)

Finally,
\[
\left\| \nabla_T \tilde{L}^{k+1,k}_{\tau,\alpha} \right\|_F = \frac{4}{\tau} \left\| \Delta^{k+1,k}_{T} \right\|_F.
\] (A.27)

Combining (A.22), (A.23), (A.25), (A.26), (A.27), we get that there exists a large enough constant \( c_0 > 0 \) independent of \( k \), such that
\[
\text{dist}(0, \partial \tilde{L}^{k+1,k}_{\tau,\alpha}) \leq c_0 \left( \left\| \Delta^{k+1,k}_{U,T} \right\|_F + \left\| \Delta^{k,k-1}_{U,T} \right\|_F \right).
\] as desired. \(\square\)

Now, we can present the proof concerning global convergence.

**Proof of Theorem 4.2** We have mentioned that \( \{ \tilde{L}^{k+1,k}_{\tau,\alpha} \} \) inherits the properties of \( \{ \tilde{L}^{k+1,k}_{\tau,\alpha} \} \), i.e., it is bounded, nonincreasing and convergent. We denote its limit as
\[
\tilde{L}^*_{\tau,\alpha} = \lim_{k \to \infty} \tilde{L}^{k+1,k}_{\tau,\alpha} = \tilde{L}^*_{\tau,\alpha}(U^*, T^*, Y^*, T^*) \text{, where } \{U^*, T^*, Y^*, T^*\} \text{ is a limit point. According to Definition 2 and Proposition A.1, there exist an } \epsilon_0 > 0, \text{ a neighborhood of } \{U^*, T^*, Y^*, T^*\}, \text{ and a continuous and concave function } \psi() : [0, \epsilon_0) \to \mathbb{R}_+ \text{ such that for all } \{U, T, Y, T'\} \in \mathcal{N} \text{ satisfying } \tilde{L}^*_{\tau,\alpha} < \tilde{L}^*_{\tau,\alpha}(U, T, Y, T') \text{, we have}
\]
\[
\psi'(\tilde{L}^*_{\tau,\alpha}(U, T, Y, T') - \tilde{L}^*_{\tau,\alpha}) \text{dist}(0, \partial \tilde{L}^*_{\tau,\alpha}(U, T, Y, T')) \geq 1. \quad (A.28)
\]

Let \( \epsilon_1 > 0 \) be such that
\[
\mathbb{R}_{\epsilon_1} := \{(U, T, Y, T') \mid \left\| U_j - U_j^* \right\|_F < \epsilon_1, 1 \leq j \leq d, \left\| T - T^* \right\|_F < \epsilon_1, \left\| Y - Y^* \right\|_F < 2\epsilon_1, \left\| T' - T^* \right\|_F < 2\epsilon_1 \} \subset \mathcal{N},
\]
and let \( \mathbb{R}_{\epsilon_1}^{U,T} := \{(U, T) \mid \left\| U_j - U_j^* \right\|_F < \epsilon_1, 1 \leq j \leq d, \left\| T - T^* \right\|_F < \epsilon_1 \} \). From the stationary point system (3.10) and the expression of \( \gamma^{k+1} \) in (3.33), we have
\[
\left\| \gamma^k - \gamma^* \right\|_F = \left\| \mathcal{W}_1^{k-1} * (T^k - A) - \mathcal{W}_1^* * (T^* - A) \right\|_F
\leq \left\| \mathcal{W}_1^{k-1} * (T^k - A) - \mathcal{W}_1^k * (T^k - A) \right\|_F
\quad + \left\| \mathcal{W}_1^k * (T^k - A) - \mathcal{W}_1^* * (T^* - A) \right\|_F
\leq \left\| \Delta^{k-1}_T \right\|_F + \left\| \Delta^*_T \right\|_F.
\] (A.29)
where the last inequality follows from Propositions 2.3 and 2.2. On the other hand,

$$
\|T^{k-1} - T^*\|_F \leq \|\Delta^{k-1}_T\|_F + \|\Delta^*_T\|_F.
$$

(A.30)

As Theorem 4.1 shows that there exists $k_0 > 0$ such that for $k \geq k_0$, $\|\Delta^{k-1}_T\|_F < \epsilon_1$, (A.29) and (A.30) tells us that if $k \geq k_0$ and $\{U^k, T^k\} \in \mathbb{B}_{\epsilon_1}^{U,T}$, then $\{U^k, T^k, \gamma^k, T^{k-1}\} \in \mathbb{B}_{\epsilon_1} \subseteq \mathcal{N}$. Such $k_0$ must exist as $\{U^*, T^*, \gamma^*, T^*\}$ is a limit point. In addition, denote $c_1 := \min\{\alpha/2, 1/\tau\}$; then, there exists $k_1 \geq k_0$ such that $\{U^{k_1}, T^{k_1}\} \in \mathbb{B}_{\epsilon_1/2}^{U,T}$ and

$$
\frac{c_0}{2\sqrt{c_1}c_2} \left\|\Delta^{k_1-1}_U, T, \gamma^k \right\|_F < \frac{\epsilon_1}{16}, \quad \frac{c_0}{2\sqrt{c_1}} \left\|\Delta^{k_1-1}_U, k_1-2 \right\|_F < \frac{\epsilon_1}{16},
$$

(A.31)

where $c_0$ is the constant appearing in Lemma A.3, and $c_2$ is a constant such that $c_2 > 16c_0/\sqrt{c_1}$.

In what follows, we use induction method to show that $\{U^k, T^k\} \in \mathbb{B}_{\epsilon_1}^{U,T}$ for all $k > k_1$. Since $\psi(\cdot)$ in Definition 2 is concave, it holds that for any $k$,

$$
\psi'(\tilde{L}^{k-1}_{\tau, \alpha} - \tilde{L}^*_\tau) \left( (\tilde{L}^{k-1}_{\tau, \alpha} - \tilde{L}^*_\tau) - (\tilde{L}^{k+1}_{\tau, \alpha} - \tilde{L}^*_\tau) \right) \leq \psi(\tilde{L}^{k-1}_{\tau, \alpha} - \tilde{L}^*_\tau) - \psi(\tilde{L}^{k+1}_{\tau, \alpha} - \tilde{L}^*_\tau); \quad (A.32)
$$

on the other side, from the previous paragraph we see that $\{U^{k_1}, T^{k_1}\} \in \mathbb{B}_{\epsilon_1/2}^{U,T}$, $\{U^{k_1}, T^{k_1}, \gamma^{k_1}, T^{k_1-1}\} \in \mathbb{B}_{\epsilon_1} \subseteq \mathcal{N}$, and so (A.28) holds at $\{U^{k_1}, T^{k_1}, \gamma^{k_1}, T^{k_1-1}\}$.

Recall $c_1 = \min\{\alpha/2, 1/\tau\}$. From Lemma A.2 and the relation between $\tilde{L}_\tau$ and $\tilde{L}_{\tau, \alpha}$, we obtain

$$
c_1 \left\|\Delta^{k_1+1}_U, T \right\|_F^2 \leq \tilde{L}^{k_1+1}_{\tau, \alpha} - \tilde{L}^{k_1+1}_{\tau, \alpha} \leq \frac{\psi(\tilde{L}^{k_1, k_1-1}_{\tau, \alpha} - \tilde{L}^*_{\tau, \alpha}) - \psi(\tilde{L}^{k_1+1, k_1-1}_{\tau, \alpha} - \tilde{L}^*_{\tau, \alpha})}{\psi(\tilde{L}^{k_1, k_1-1}_{\tau, \alpha} - \tilde{L}^*_{\tau, \alpha})} \leq c_2 \left( \psi(\tilde{L}^{k_1+1, k_1-1}_{\tau, \alpha} - \tilde{L}^*_{\tau, \alpha}) \right) \cdot \frac{c_2^2}{\epsilon_1^2} \cdot \text{dist}(0, \partial \tilde{L}^{k_1, k_1-1}_{\tau, \alpha}),
$$

where the second inequality is due to (A.32) while the last one comes from (A.28). Using $\sqrt{ab} \leq \frac{a+b}{2}$ for $a \geq 0, b \geq 0$, invoking (A.19) and noticing the range in (A.31), we obtain

$$
\sqrt{c_1} \left\|\Delta^{k_1+1}_U, T \right\|_F \leq \frac{c_2}{2} \left( \psi(\tilde{L}^{k_1, k_1-1}_{\tau, \alpha} - \tilde{L}^*_{\tau, \alpha}) - \psi(\tilde{L}^{k_1+1, k_1-1}_{\tau, \alpha} - \tilde{L}^*_{\tau, \alpha}) \right) + \frac{c_0}{2c_2} \left( \left\|\Delta^{k_1, k_1-2}_U, T \right\|_F + \left\|\Delta^{k_1-1, k_1-2}_U, T \right\|_F \right),
$$

and so

$$
\left\|\Delta^{k_1+1}_U, T \right\|_F \leq \left\|\Delta^{k_1+1}_U \right\|_F + \left\|\Delta^{k_1}_U, T \right\|_F \leq \frac{\epsilon_1}{2} + \frac{\epsilon_1}{2} = \epsilon_1,
$$

\text{ Springer}
namely, \( \{U^{k_1 + 1}, T^{k_1 + 1}\} \in \mathbb{B}_{e_1} U, T \).

Now, assume that \( \{U^k, T^k\} \in \mathbb{B}_{e_1} U, T \) for \( k = k_1, \ldots, K \). This implies that (A.28) is true at \( \{U^k, T^k, \gamma^k, T^{k-1}\} \), and similarly to the above analysis, we have

\[
\sqrt{c_1} \left\| \Delta_{U, T}^{k+1, k} \right\| F \leq \frac{c_2}{2} \left( \psi(\tilde{L}_{r, \alpha}^{k, k-1} - \tilde{L}_{r, \alpha}^*) - \psi(\tilde{L}_{r, \alpha}^{k+1, k} - \tilde{L}_{r, \alpha}^*) \right) + \frac{c_0}{2c_2} \left( \left\| \Delta_{U, T}^{k, k-1} \right\| F + \left\| \Delta_{U, T}^{k-1, k-2} \right\| F \right), \quad k = k_1, \ldots, K.
\] (A.33)

We then show that \( \{U^{K+1}, T^{K+1}\} \in \mathbb{B}_{e_1} U, T \). Summing (A.33) for \( k = k_1, \ldots, K \) yields

\[
\sqrt{c_1} \sum_{k=k_1}^{K} \left\| \Delta_{U, T}^{k+1, k} \right\| F \leq \frac{c_2}{2} \left( \psi(\tilde{L}_{r, \alpha}^{k_1, k_1-1} - \tilde{L}_{r, \alpha}^*) - \psi(\tilde{L}_{r, \alpha}^{K+1, K} - \tilde{L}_{r, \alpha}^*) \right) + \frac{c_0}{2c_2} \sum_{k=k_1}^{K} \left( \left\| \Delta_{U, T}^{k, k-1} \right\| F + \left\| \Delta_{U, T}^{k-1, k-2} \right\| F \right)
\]
\[
\leq \frac{c_2}{2} \left( \psi(\tilde{L}_{r, \alpha}^{k_1, k_1-1} - \tilde{L}_{r, \alpha}^*) - \psi(\tilde{L}_{r, \alpha}^{K+1, K} - \tilde{L}_{r, \alpha}^*) \right)
\]
\[
+ \frac{c_0}{c_2} \sum_{k=k_1}^{K} \left\| \Delta_{U, T}^{k+1, k} \right\| F + \frac{2c_0}{c_2} \left\| \Delta_{U, T}^{k_1, k_1-1} \right\| F + \frac{c_0}{c_2} \left\| \Delta_{U, T}^{k_1-1, k_1-2} \right\| F.
\] (A.34)

Rearranging the terms, noticing (A.31) and noticing that \( \frac{c_2}{c_0} > \frac{\sqrt{c_1}}{16} \), we have

\[
\frac{15\sqrt{c_1}}{16} \sum_{k=k_1}^{K} \left\| \Delta_{U, T}^{k+1, k} \right\| F \leq \frac{\sqrt{c_1}}{4} e_1 + \frac{\sqrt{c_1} e_1}{16} + \frac{\sqrt{c_1} e_1}{16},
\]

and so

\[
\left\| \Delta_{U, T}^{K+1, *} \right\| F \leq \left\| \Delta_{U, T}^{K+1, K_1} \right\| F + \left\| \Delta_{U, T}^{K_1, *} \right\| F
\]
\[
< \sum_{k=k_1}^{K} \left\| \Delta_{U, T}^{k+1, k} \right\| F + \frac{e_1}{2}
\]
\[
< \frac{3e_1}{8} + \frac{e_1}{2} < e_1.
\]

Thus, induction method implies that \( \{U^k, T^k\} \in \mathbb{B}_{e_1} U, T \) for all \( k \geq k_1 \), i.e., \( \{U^k, T^k, \gamma^k, T^{k-1}\} \in \mathcal{N}, k \geq k_1 \). As a result, (A.33) holds for all \( k \geq k_1 \), so does (A.34). Therefore, letting \( K \to \infty \) in (A.34) yields

\[
\sum_{k=1}^{\infty} \left\| \Delta_{U, T}^{k+1, k} \right\| F < +\infty.
\]
which shows that \( \{U^k, T^k\} \) is a Cauchy sequence and hence converges. Since \( \{U^*, T^*\} \) in Theorem 4.1 is a limit point, the whole sequence converges to \( \{U^*, T^*\} \). This completes the proof.

**Acknowledgements** We thank the editor and the anonymous reviewers for their insightful comments and suggestions that helped improve this manuscript.

**Funding** The first author was supported by the National Natural Science Foundation of China Grants 11801100 and 12171105, and the Fok Ying Tong Education Foundation Grant 171094. The second author was supported by the Simons Foundation Collaboration Grant 572064.

**Declarations**

**Conflict of interest** The authors declare no competing interests.

**References**

1. Anandkumar, A., Jain, P., Shi, Y., Niranjan, U.N.: Tensor vs. matrix methods: robust tensor decomposition under block sparse perturbations. In: Artificial Intelligence and Statistics, pp. 268–276 (2016)
2. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Math. Program. 137(1-2), 91–129 (2013)
3. Beaton, A., Tukey, J.: The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 16(2), 147–185 (1974)
4. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. 146(1-2), 459–494 (2014)
5. Chen, J., Saad, Y.: On the tensor SVD and the optimal low rank orthogonal approximation of tensors. SIAM. J. Matrix Anal. Appl. 30(4), 1709–1734 (2009)
6. Cheng, L., Wu, Y.C., Poor, H.V.: Probabilistic tensor canonical polyadic decomposition with orthogonal factors. IEEE Trans. Signal Process. 65(3), 663–676 (2016)
7. Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Ciaia, C., Phan, H.A.: Tensor decompositions for signal processing applications: from two-way to multiway component analysis. IEEE Signal Process. Mag. 32(2), 145–163 (2015)
8. De Almeida, A.L.F., Kibangou, A.Y., Miron, S., Araújo, D.C.: Joint data and connection topology recovery in collaborative wireless sensor networks. In: Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 5303–5307 (2013)
9. De Lathauwer, L.: Algebraic methods after prewhitening. In: Handbook of Blind Source Separation, pp. 155–177. Elsevier (2010)
10. De Lathauwer, L.: A Short introduction to tensor-based methods for factor analysis and blind source separation. In: Proceeding of the IEEE International Symposium on Image and Signal Processing and Analysis (ISPA 2011), pp. 558–563 (2011)
11. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21, 1253–1278 (2000)
12. Ding, M., Huang, T.Z., Ma, T.H., Zhao, X.L., Yang, J.H.: Cauchy noise removal using group-based low-rank prior. Appl. Math. Comput. 372, 124971 (2020)
13. Feng, Y., Fan, J., Suykens, J.: A statistical learning approach to modal regression. J. Mach. Learn. Res. 21(2), 1–35 (2020)
14. Feng, Y., Huang, X., Shi, L., Yang, Y., Suykens, J.: Learning with the maximum correntropy criterion induced losses for regression. J. Mach. Learn. Res. 16, 993–1034 (2015)
15. Ganan, S., McClure, D.: Bayesian image analysis: an application to single photon emission tomography. Amer. Statist. Assoc. 12–18 (1985)
16. Goldfarb, D., Qin, Z.: Robust low-rank tensor recovery: models and algorithms. SIAM J. Matrix Anal. Appl. 35(1), 225–253 (2014)
17. Guan, N., Liu, T., Zhang, Y., Tao, D., Davis, L.S.: Truncated cauchy non-negative matrix factorization. IEEE Trans. Pattern Anal. Mach. Intell. 41(1), 246–259 (2017)
18. Guan, Y., Chu, D.: Numerical computation for orthogonal low-rank approximation of tensors. SIAM J. Matrix Anal. Appl. 40(3), 1047–1065 (2019)
19. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 33(8), 1561–1576 (2010)
20. Hillar, C.J., Lim, L.H.: Most tensor problems are NP-hard. J. ACM 60(6), 45:1–45:39 (2013)
21. Holland, P., Welsch, R.: Robust regression using iteratively reweighted least-squares. Commun. Stat.-Theory Methods 6(9), 813–827 (1977)
22. Hong, D., Kolda, T.G., Duersch, J.A.: Generalized canonical polyadic tensor decomposition. SIAMRev. 62(1), 133–163 (2020)
23. Hong, M., Luo, Z.Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. 26(1), 337–364 (2016)
24. Hu, S., Li, G.: Convergence rate analysis for the higher order power method in best rank one approximations of tensors. Numer. Math. 140(4), 993–1031 (2018)
25. Hu, S., Ye, K. (2019)
26. Huber, P.J.: Robust statistics, vol. 523. Wiley, New York (2004)
27. Kim, G., Cho, J., Kang, M.: Cauchy noise removal by weighted nuclear norm minimization. J. Sci. Comput. 83, 15 (2020)
28. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. 51, 455–500 (2009)
29. Kovnatsky, A., Glashoff, K., Bronstein, M.M.: MADMM: a generic algorithm for non-smooth optimization on manifolds. In: European Conference on Computer Vision, pp. 680–696. Springer (2016)
30. Li, G., Liu, T., Pong, T.K.: Peaceman–Rachford splitting for a class of nonconvex optimization problems. Comput. Optim. Appl. 68(2), 407–436 (2017)
31. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. 25(4), 2434–2460 (2015)
32. Li, G., Pong, T.K.: Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Math. Program. 159(1-2), 371–401 (2016)
33. Li, G., Pong, T.K.: Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. Found. Comput. Math. 18(5), 1199–1232 (2018)
34. Li, J., Usevich, K., Comon, P.: Globally convergent Jacobi-type algorithms for simultaneous orthogonal symmetric tensor diagonalization. SIAMJ. Matrix Anal. Appl. 39(1), 1–22 (2018)
35. Li, J., Zhang, S.: Polar decomposition based algorithms on the product of stiefel manifolds with applications in tensor approximation. arXiv:1912.10390 (2019)
36. Li, X., Lu, Q., Dong, Y., Tao, D.: Robust subspace clustering by cauchy loss function. IEEE Trans. Neural Netw. Learn. Syst. 30(7), 2067–2078 (2018)
37. Liu, H., So, A.M.C., Wu, W.: Quadratic optimization with orthogonality constraint: explicit Lojasiewicz exponent and linear convergence of retraction-based line-search and stochastic variance-reduced gradient methods. Math. Program. 178(1), 215–262 (2019)
38. Maronna, R., Bustos, O., Yohai, V.: Bias-and efficiency-robustness of general M-estimators for regression with random carriers. In: Smoothing Techniques for Curve Estimation, pp. 91–116. Springer (1979)
39. Mei, J.J., Dong, Y., Huang, T.Z., Yin, W.: Cauchy noise removal by nonconvex admm with convergence guarantees. J. Sci. Comput. 74(2), 743–766 (2018)
40. Pan, J., Ng, M.K.: Symmetric orthogonal approximation to symmetric tensors with applications to image reconstruction. Numer. Linear Algebra Appl. 25(5), e2180 (2018)
41. Pravdova, V., Estienne, F., Walczak, B., Massart, D.L.: A robust version of the Tucker3 model. Chemometr. Intell. Lab. Syst. 59(1), 75–88 (2001)
42. Savas, B., Lim, L.H.: Quasi-Newton methods on grassmannians and multilinear approximations of tensors. SIAM J. Sci. Comput. 32(6), 3352–3393 (2010)
43. Sciacchitano, F., Dong, Y., Zeng, T.: Variational approach for restoring blurred images with cauchy noise. SIAM J. Imag. Sci. 8(3), 1894–1922 (2015)
44. Shashua, A., Levin, A.: Linear image coding for regression and classification using the tensor-rank principle. In: CVPR, vol. 1, pp. 1–I. IEEE (2001)
45. Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C.: Tensor decomposition for signal processing and machine learning. IEEE Trans. Signal Process. 65(13), 3551–3582 (2017)
46. Sidiropoulos, N.D., Giannakis, G.B., Bro, R.: Blind parafac receivers for ds-cdma systems. IEEE Trans. Signal Process. 48(3), 810–823 (2000)
47. Signoretto, M., Dinh, Q.T., De Lathauwer, L., Suykens, J.A.K.: Learning with tensors: a framework based on convex optimization and spectral regularization. Mach. Learn. 94(3), 303–351 (2014)
48. Sørensen, M., De Lathauwer, L., Comon, P., Icart, S., Deneire, L.: Canonical polyadic decomposition with a columnwise orthonormal factor matrix. SIAM J. Matrix Anal. Appl. 33(4), 1190–1213 (2012)
49. Sørensen, M., De Lathauwer, L., Deneire, L.: PARAFAC with orthogonality in one mode and applications in DS-CDMA systems. In: Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), pp. 4142–4145 (2010)
50. Vervliet, N., Debals, O., Sorber, L., Van Barel, M., De Lathauwer, L.: Tensorlab 3.0. http://www.tensorlab.net. Available online (2016)
51. Wang, L., Chu, M.T., Yu, B.: Orthogonal low rank tensor approximation: alternating least squares method and its global convergence. SIAM J. Matrix Anal. and Appl. 36(1), 1–19 (2015)
52. Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Sci. Comput. 78(1), 29–63 (2019)
53. Yang, Y.: The epsilon-alternating least squares for orthogonal low-rank tensor approximation and its global convergence. SIAM J. Matrix Anal. Appl. 41(4), 1797–1825 (2020)
54. Yang, Y., Feng, Y., Suykens, J.A.K.: Robust low-rank tensor recovery with regularized redescending m-estimator. IEEE Trans. Neural Netw. Learn. Syst. 27(9), 1933–1946 (2015)
55. Ye, K., Hu, S.: When geometry meets optimization theory: partially orthogonal tensors. arXiv:2201.04824 (2022)
56. Yu, P., Li, G., Pong, T.K.: Kurdyka–Łojasiewicz exponent via inf-projection. Found. Comput. Math. 1–47 (2021)

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.