Numerische Mathematik (2023) 153:141–170
https://doi.org/10.1007/s00211-022-01328-6

Numerische Mathematik

Nonnegative low rank tensor approximations with multidimensional image applications

Tai-Xiang Jiang[1,2] · Michael K. Ng[3] · Junjun Pan[3] · Guang-Jing Song[4]

Received: 26 December 2020 / Revised: 7 October 2022 / Accepted: 7 October 2022 /
Published online: 29 October 2022

Abstract

The main aim of this paper is to develop a new algorithm for computing a nonnegative low rank tensor approximation for nonnegative tensors that arise in many multidimensional imaging applications. Nonnegativity is one of the important properties, as each pixel value refers to a nonzero light intensity in image data acquisitions. Our approach is different from classical nonnegative tensor factorization (NTF), which requires each factorized matrix, and/or tensor, to be nonnegative. In this paper, we determine a nonnegative low Tucker rank tensor to approximate a given nonnegative tensor. We propose an alternating projections algorithm for computing such a nonnegative low rank tensor approximation, which is referred to as NLRT. The convergence of the proposed manifold projection method is established. The experimental results for synthetic data and multidimensional images are presented to demonstrate that the performance of NLRT is better than the state-of-the-art NTF methods.

Mathematics Subject Classification 15A69 · 15A72 · 65Y20

✉ Guang-Jing Song
sgjshu@163.com

Tai-Xiang Jiang
taixiangjiang@gmail.com

Michael K. Ng
mng@maths.hku.hk

Junjun Pan
junjpan@hku.hk

1 School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, People's Republic of China

2 Kash Institute of Electronics and Information Industry, Kashi, People's Republic of China

3 Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong, People's Republic of China

4 School of Mathematics and Information Sciences, Weifang University, Weifang 261061, People's Republic of China

1 Introduction

Nonnegative data are very common in many data analysis applications. For instance, in image analyses, image pixel values are nonnegative, and the associated images can be seen as nonnegative matrices for clustering and recognition tasks. When the data are already high dimensional by nature, for example, video data, hyperspectral data, fMRI data and so on, it then seems more natural to represent the information in a high dimensional space, rather than flatten the data into a matrix. The data represented in high dimensions are referred to as tensors.

An m-dimensional tensor $\mathcal{A}$ is a multidimensional array, $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_m}$. To extract pertinent information from given large tensor data, low rank tensor decompositions are usually considered. In recent decades, various tensor decompositions have been developed according to different applications. The most famous and widely used decompositions are the canonical polyadic decomposition (CPD) and Tucker decomposition. For more details of tensor applications and tensor decompositions, we refer to the review papers [13, 24]. In this paper, we only target tensors in Tucker form. Hence, in the following, we briefly review the Tucker decomposition.

Given a tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m}$, the Tucker decomposition [8, 13, 26] is defined as follows:

$$\mathcal{A} = \mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \cdots \times_m \mathbf{U}^{(m)}, \tag{1}$$

that is,

$$\mathcal{A}_{i_1,\cdots,i_m} = \sum_{j_1,\cdots,j_m} \mathcal{G}_{j_1,\cdots,j_m} \mathbf{U}^{(1)}_{i_1,j_1} \cdots \mathbf{U}^{(m)}_{i_m,j_m}, \tag{2}$$

where $\mathcal{G} = (\mathcal{G}_{j_1,j_2,\cdots,j_m}) \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_m}$, $\mathbf{U}^{(k)}$ is an n_k-by-r_k matrix (whose columns are usually mutually orthogonal), and $\times_k$ denotes the k-mode matrix product of a tensor defined by

$$(\mathcal{G} \times_k \mathbf{U}^{(k)})_{j_1 \cdots j_{k-1} i_k j_{k+1} \cdots j_m} = \sum_{j_k=1}^{r_k} \mathcal{G}_{j_1 \cdots j_{k-1} j_k j_{k+1} \cdots j_m} \mathbf{U}^{(k)}_{i_k,j_k}.$$

The minimal value of $(r_1, r_2, \ldots, r_m)$ is defined as the Tucker (or multilinear) rank of $\mathcal{A}$, denoted as $\text{rank}_T(\mathcal{A}) = (r_1, r_2, \cdots, r_m)$.

Since high-dimensional nonnegative data are everywhere in the real world, and the nonnegativity of the factor matrices derived from tensor decompositions can lead to interpretations for real applications, many nonnegative tensor decompositions have been proposed and developed, and most of them are based on a tensor decomposition with nonnegative constraints. A Tucker decomposition with nonnegative constraints, which is referred to as a nonnegative Tucker decomposition (NTD) in [12], aims to solve

$$\begin{aligned} &\min \|\mathcal{A} - \mathcal{X}\|_F^2 \\ &\text{s.t.} \quad \mathcal{X} = \mathcal{S} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \times_3 \cdots \times_m \mathbf{P}_m, \\ &\qquad \mathcal{S} \in \mathbb{R}_+^{r_1 \times \cdots \times r_m}, \quad \mathbf{P}_k \in \mathbb{R}_+^{n_k \times r_k}, \quad k = 1, \ldots m, \end{aligned} \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a tensor (see the definition in Sect. 2), and $\mathbb{R}_+^{s_1\times\cdots\times s_m}$ is the set of nonnegative $s_1\times\cdots\times s_m$ tensors (or matrices) whose entries are nonnegative. In [12], Kim and Choi first studied this model and proposed multiplicative updating (MU) algorithms extended from nonnegative matrix factorization (NMF) to solve it. In [32], Zhou et al. transformed this problem into a series of NMF problems and used the MU and hierarchical alternating least squares (HALS) algorithms on the unfolding matrices for Tucker decomposition calculations. Other constraints, such as the orthogonality of the factor matrices, are also considered and studied by some researchers [19, 21]. For instance, in [21], Pan et al. proposed an orthogonal nonnegative Tucker decomposition and applied the alternating direction method of multipliers (ADMM) to obtain the clustering information from the factor matrices and the joint connection weight from the core tensor.

The greatest advantage of the NTD model is that the core tensor and factor matrices can be interpretable thanks to the requirement of the factorized components. However, the approximation $\mathcal{X}$ is not the best approximation of $\mathcal{A}$ for the given Tucker rank $(r_1,\ldots,r_m)$. Hence, it is required to find the best low Tucker rank nonnegative approximation for a given nonnegative tensor $\mathcal{A}$ with interpretable factor matrices and a core tensor. In this paper, we propose the following problem. Given a nonnegative tensor $\mathcal{A}\in\mathbb{R}_+^{n_1\times\cdots\times n_m}$, we consider

$$\min_{\mathcal{X}\geq 0}\|\mathcal{A}-\mathcal{X}\|_F^2,\quad \text{s.t.}\quad rank_T(\mathcal{X})=(r_1,r_2,\ldots,r_m). \tag{4}$$

From $rank_T(\mathcal{X})=(r_1,r_2,\ldots,r_m)$, we can deduce that there exist a core tensor $\mathcal{S}\in\mathbb{R}^{r_1\times r_2\times\cdots\times r_m}$ and orthogonal factor matrices $\{\mathbf{P}_k:\mathbf{P}_k\in\mathbb{R}^{n_k\times r_k},\mathbf{P}_k^T\mathbf{P}_k=\mathbf{I}_{r_k},k=1,\ldots,m\}$, such that

$$\mathcal{X}=\mathcal{S}\times_1\mathbf{P}_1\times_2\mathbf{P}_2\times_3\cdots\times_m\mathbf{P}_m,$$

and the entries of $\mathcal{X}$ are nonnegative ($\mathcal{X}\geq 0$). For $k=1,\ldots,m$, let $\mathbf{X}_k$ be the k-th unfolding of the tensor $\mathcal{X}$, defined as $\mathbf{X}_k\in\mathbb{R}^{n_k\times(n_{k+1}\cdots n_m n_1\cdots n_{k-1})}$. From the definition of the Tucker decomposition, we deduce that $r_k=rank(\mathbf{X}_k)$, and the factor matrix $\mathbf{P}_k$ can be obtained by a singular value decomposition of $\mathbf{X}_k$:

$$\mathbf{X}_k=\mathbf{P}_k\mathbf{\Sigma}_k\mathbf{Q}_k^T,$$

Here, $\mathbf{\Sigma}_k$ is a diagonal matrix of size r_k-by-r_k, and $\mathbf{Q}_k$ is $\prod_{i\neq k}^m n_i$-by-r_k with orthonormal columns ($\mathbf{Q}_k^T$ is the transpose of $\mathbf{Q}_k$).

We remark that problem (4) without the nonnegativity constraint on the approximation $\mathcal{X}$ is referred to as the best low multilinear rank approximation problem, which has been well discussed and used widely as a tool in dimensionality reduction and signal subspace estimation during the last two decades. The classical methods for the problem are the truncated higher-order SVD (HOSVD) [8] and the higher-order orthogonal iteration (HOOI) [9, 14], which is a higher-order extension of an iteration method for matrices. Without the nonnegative constraint, the solution $\mathcal{X}$ may have neg-

ative entries that cannot preserve the nonnegative property of the given nonnegative tensor.

Note that in the proposed model (4), we require $\mathcal{X}$ to be nonnegative, while its factorized components $(\mathcal{S}, \{\mathbf{P}_k\}_{k=1}^m)$ are not necessarily nonnegative. For example, given a hyperspectral image $\mathcal{A}$, $\mathcal{X}$ can be seen as an approximate image of $\mathcal{A}$, but with a lower multilinear rank. On the one hand, we keep the approximate image $\mathcal{X}$ as nonnegative. On the other hand, no constraints are added to the factorized components. Therefore we can consider a similar idea utilized in HOSVD to identify important features in the approximation, which are ranked based on their importance. Then we can identify the important factorized components for classification purposes; see Sect. 4.5 for an example.

1.1 Outline and contributions

The main aim of this paper is to propose and study low multiliear rank nonnegative tensor approximations for multidimensional image applications. In Sect. 2, we propose an alternating manifold projection method for computing the nonnegative low multilinear rank tensor approximation. The projection method is developed by constructing two projections; one is a combination of a projection of low rank matrix manifolds and the nonnegative projection; the other one is a projection of taking the average of the tensors. In Sect. 3, the convergence of the proposed method is studied and shown. Section 4 presents the experimental results for synthetic data and multidimensional images in noisy and noise-free cases. It demonstrates that the performance of the proposed nonnegative low multilinear rank tensor approximation method is better than the state-of-the-art NTF methods. Some concluding remarks are given in Sect. 5.

2 Nonnegative low rank tensor approximation

Let us first start with some tensor operations used throughout this paper. The inner product of two same-sized tensors $\mathcal{A}$ and $\mathcal{B}$ is defined as

$$\langle \mathcal{A}, \mathcal{B} \rangle := \sum_{i_1, i_2, \cdots, i_m} \mathcal{A}_{i_1 i_2 \cdots i_m} \mathcal{B}_{i_1 i_2 \cdots i_m}.$$

The Frobenius norm of an m-dimensional tensor $\mathcal{A}$ is defined as

$$\|\mathcal{A}\|_F := \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle} = \left(\sum_{i_1, i_2, \cdots, i_m} \mathcal{A}^2_{i_1 i_2 \cdots i_m} \right)^{\frac{1}{2}}.$$

2.1 The optimization model

We first give the following lemma to demonstrate that the set of constraints in (4) is nonempty.

Lemma 1 *The set of constraints* $\{\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_m} \mid \text{rank}(\boldsymbol{X}_k) = r_k \ (k = 1..., m), \mathcal{X} \geq 0\}$ *in (4) is nonempty.*

Proof First, we will prove that there always exists a tensor $\mathcal{S} \in \mathbb{R}_+^{r_1 \times \cdots \times r_m}$ that has full unfolding matrix rank for each mode.

For any $t \in \mathbb{R}_+^{r_1 r_2 \cdots r_m}$, let $(\mathbf{S}_k)(t) \in \mathbb{R}^{r_k \times r_1 \cdots r_{k-1} r_{k+1} \cdots r_m}$ hold the elements of t. Let $(\mathbf{S}_k)(t)_{r_k}$ be the $r_k \times r_k$ sub matrix of $(\mathbf{S}_k)(t)$ and $det((\mathbf{S}_k)(t)_{r_k})$ be its determinant. As we know that $det((\mathbf{S}_k)(t)_{r_k})$ is a polynomial in the entries of t, it either vanishes on a set of zero measures or it is a zero polynomial. We may choose $(\mathbf{S}_k)(t)_{r_k}$ to be the identity matrix, which implies that $det((\mathbf{S}_k)(t)_{r_k})$ are not zero polynomials. This means the Lebesgue measure of the space whose $det((\mathbf{S}_k)(t)_{r_k}) = 0$ is zero, i.e., the rank of $(\mathbf{S}_k)(t)_{r_k}$ is r_k almost everywhere.

Thus, for $k = 1, \ldots, m$, construct $\mathcal{T}_k = \{\mathcal{S} \in \mathbb{R}_+^{r_1 \times \cdots \times r_m} | rank(\mathcal{S}_k) = r_k\}$, and let $\bar{\mathcal{T}}_k$ be its complement. From the above analysis, we know that the Lebesgue measure of $\bar{\mathcal{T}}_k$ is equal to zero. Let $\mathcal{T} = \cap_{k=1}^m \mathcal{T}_k$; then, its complement $\bar{\mathcal{T}} = \cup_{k=1}^m \bar{\mathcal{T}}_k$, and its Lebesgue measure is the summation of that of $\bar{\mathcal{T}}_k$ from $k = 1$ to $k = m$, equal to zero. This implies that the Lebesque measure of $\mathcal{T}$ equals 1, i.e., $\mathcal{S} \in \mathbb{R}_+^{r_1 \times \cdots \times r_m}$ of unfolding matrix rank $(r_1, \ldots, r_m)$ exists almost everywhere.

Suppose $\mathbf{P}_k \in \mathbb{R}^{n_k \times r_k}$, and $\mathbf{P}_k = [\mathbf{I}_k | \mathbf{U}_k]$, where $\mathbf{I}_k$ is the identity matrix of r_k and $\mathbf{U}_k \in \mathbb{R}^{r_k \times (n_k - r_k)}$ is a random nonnegative matrix for all $k = 1, \ldots, m$. Construct

$$\mathcal{X} = \mathcal{S} \times \mathbf{P}_1 \times \cdots \times \mathbf{P}_m,$$

we obtain that $\mathcal{X}$ is nonnegative and its multilinear rank is $(r_1, \cdots, r_m)$. Hence the set of constraints is nonempty. □

From the definition of Tucker decomposition and the property of multilinear rank that $r_k = rank(\mathbf{X}_k)$ for $k = 1, \cdots, m$, the mathematical model in (4) can be reformulated as the following optimization problem:

$$\min_{\substack{\text{rank}(\mathbf{X}_k) = r_k, \mathbf{X}_k \geq 0 \\ (k=1,\ldots,m)}} \sum_{k=1}^{m} \|\mathbf{A}_k - \mathbf{X}_k\|_F^2, \tag{5}$$

where $\mathbf{X}_k$ and $\mathbf{A}_k$ are the k-th modes of the unfolding matrix of $\mathcal{X}$ and $\mathcal{A}$, respectively. The sizes of $\mathbf{A}_k$ and $\mathbf{X}_k$ are n_k-by-N_k with $N_k = \prod_{i \neq k}^m n_i$.

Note that from (5), $\{\mathbf{X}_k\}_{k=1}^m$ can be seen as m manifolds of low rank and nonnegative matrices. Meanwhile, as the Frobenius norm is employed in the objective function, to a certain extent, our model is tolerant to noise, which is unavoidable in real-world data. In the next section, an alternating projections on the manifolds algorithm will be proposed to solve model in (5).

2.2 The proposed algorithm

To start showing the proposed algorithm for (5), we first define two projections. Let

$$\mathtt{M} = \{\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_m} \mid \mathcal{X}_{i_1 i_2 \cdots i_m} \geq 0\} \tag{6}$$

be the set of nonnegative tensors. Then the nonnegative projection that projects a given tensor onto the tensor manifold $\mathtt{M}$ can be expressed as follows:

$$\pi(\mathcal{X}) = \begin{cases} \mathcal{X}_{i_1 i_2 \cdots i_m}, & \text{if } \mathcal{X}_{i_1 i_2 \cdots i_m} \geq 0, \\ 0, & \text{if } \mathcal{X}_{i_1 i_2 \cdots i_m} < 0. \end{cases} \tag{7}$$

Let

$$\mathtt{M}_k = \{\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_m} \mid \operatorname{rank}(\mathbf{X}_k) = r_k \}, \ k = 1, ..., m, \tag{8}$$

be the set of tensors whose k-mode unfolding matrices have fixed rank r_k. By the Eckart-Young-Mirsky theorem [11], the k-mode projections that project tensor $\mathcal{X}$ onto $\mathtt{M}_k$ are presented as follows:

$$\pi_k(\mathcal{X}) = \operatorname{fold}_k \left(\sum_{i=1}^{r_i} \sigma_i(\mathbf{X}_k) u_i(\mathbf{X}_k) v_i(\mathbf{X}_k)^T \right), \quad k = 1, ..., m, \tag{9}$$

where $\mathbf{X}_k$ is the k-mode unfolding matrix of $\mathcal{X}$, $\sigma_i(\mathbf{X}_k)$ is the i-th singular value of $\mathbf{X}_k$, and their corresponding left and right singular vectors are $u_i(\mathbf{X}_k)$ and $v_i(\mathbf{X}_k)$, respectively. "fold_k" denotes the operator that folds a matrix into a tensor along the k-mode.

In model (5), the multilinear rank of the nonnegative approximation $\mathcal{X}$ is required to be $(r_1, \ldots, r_m)$, which means $\mathcal{X}$ will fall in the intersection of the sets $\{\mathtt{M}_k\}_{k=1}^m$ and the nonnegative tensor set $\mathtt{M}$, i.e., $\mathcal{X} \in \bigcap_{k=1}^m (\mathtt{M}_k \bigcap \mathtt{M})$. In the following, we define two tensor sets on the product space $\mathbb{R}^{n_1 \times \cdots \times n_m} \times \cdots \times \mathbb{R}^{n_1 \times \cdots \times n_m}$ (m times of $\mathbb{R}^{n_1 \times \cdots \times n_m}$) and their corresponding projections:

•

$$\Omega_1 = \{(\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_m) : \mathcal{X}_1 = \mathcal{X}_2 = \cdots = \mathcal{X}_m \in \mathtt{M}\}. \tag{10}$$

We remark that Ω_1 is a convex and affine manifold since $\mathtt{M}$ is a convex set and an affine manifold. The projection π_{Ω_1} defined on Ω_1 is given by

$$\begin{aligned} &\pi_{\Omega_1}(\mathcal{X}_1, \cdots, \mathcal{X}_m) \\ &\quad = \left(\frac{1}{m} (\pi(\mathcal{X}_1) + \cdots + \pi(\mathcal{X}_m)), \cdots, \frac{1}{m} (\pi(\mathcal{X}_1) + \cdots + \pi(\mathcal{X}_m)) \right), \end{aligned} \tag{11}$$

where π is defined in (7).

•

$$\Omega_2 = \{(\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_m) : \mathcal{X}_1 \in \mathtt{M}_1, \mathcal{X}_2 \in \mathtt{M}_2, \ldots, \mathcal{X}_m \in \mathtt{M}_m\}. \tag{12}$$

For each $i \in \{1, ..., m\}$, $\mathtt{M}_i$ is a C^∞ manifold (Example 2 in [17]). Hence, Ω_2 can be regarded as a product of m C^∞ manifolds, i.e., $\Omega_2 = \mathtt{M}_1 \times \mathtt{M}_2 \cdots \times \mathtt{M}_m$. The projection π_{Ω_2} on Ω_2 is given by

$$\pi_{\Omega_2}(\mathcal{X}) = (\pi_1(\mathcal{X}), \ldots, \pi_m(\mathcal{X})), \tag{13}$$

where π_k ($k = 1, ..., m$) are defined in (9).

We alternately project the given $\mathcal{A}$ onto Ω_1 and Ω_2 by the projections $\pi_{\Omega_1}(\mathcal{X})$ and $\pi_{\Omega_2}(\mathcal{X})$ until it is convergent, and refer the algorithm to as the alternating projections

algorithm for the nonnegative low rank tensor (NLRT) approximation problem. The proposed algorithm is summarized in Algorithm 1. Note that the dominant overall computational cost of Algorithm 1 can be expressed as the SVDs of m unfolding matrices with sizes n_k by $N_k = \prod_{i\neq k}^{n} n_j$, respectively, which leads to a total of $O((\prod_{j=1}^{m} n_j)\sum_{i=1}^{m} r_i)$ flops.

Algorithm 1 Alternating Projections Algorithm for Nonnegative Low Rank Tensor (NLRT) Approximation

Input: Given a nonnegative tensor $\mathcal{A} \in \mathbb{R}^{n_1\times\cdots\times n_m}$, this algorithm computes a Tucker rank $(r_1, r_2, ..., r_m)$ nonnegative tensor close to A with respect to (5).

1: Initialize $s = 0$, $\mathcal{Z}_1^{(0)} = ... = Z_m^{(0)} = \mathcal{A}$ and $Z^{(0)} = (\mathcal{Z}_1^{(0)}, \mathcal{Z}_2^{(0)}, ..., \mathcal{Z}_m^{(0)})$
2: **while** the convergence criterion is not satisfied
3: $s = s + 1$
4: $(\mathcal{Y}_1^{(s)}, \mathcal{Y}_2^{(s)}, ..., \mathcal{Y}_m^{(s)}) = \pi_{\cdot\cdot 2}(\mathcal{Z}_1^{(s-1)}, \mathcal{Z}_2^{(s-1)}, \cdots, \mathcal{Z}_m^{(s-1)})$;
5: $(\mathcal{Z}_1^{(s)}, \mathcal{Z}_2^{(s)}, \cdots, \mathcal{Z}_m^{(s)}) = \pi_{\cdot\cdot 1}(\mathcal{Y}_1^{(s)}, \mathcal{Y}_2^{(s)}, ..., \mathcal{Y}_m^{(s)})$;
6: **end while**

Output: $Z^{(s)} = (\mathcal{Z}_1^{(s)}, \mathcal{Z}_2^{(s)}, \cdots, \mathcal{Z}_m^{(s)})$

Remark: The convergence criterion can be used by setting the maximum number of iterations; or/and the relative difference between successive iterates $\|Z^{(s)} - Z^{(s-1)}\|_F / \|Z^{(s-1)}\|_F$ being less than a positive number ϵ. In our numerical results, we set the convergence criterion based on the relative difference with $\epsilon = 10^{-5}$.

3 Convergence analysis

The framework of this algorithm is the same as the convex problem of finding a point in the intersection of several closed sets, and the projection sets here are two product manifolds. In [17], Lewis and Malick proved that a sequence of alternating projections converges locally linearly if the two projected sets are C^2-manifolds intersecting transversally. Lewis et al. [15] proved local linear convergence when two projected sets intersect nontangentially in the sense of linear regularity, and one of the sets is super regular. Later, Bauschke et al. [3, 4] further investigated the case of nontangential intersections and proved linear convergence under weaker regularity and the transversality hypotheses. In [20], Noll and Rondepierre generalized the existing results by studying the intersection condition of the two projected sets. They established local convergence of alternating projections between subanalytic sets under a mild regularity hypothesis on one of the sets. Here, we analyze the convergence of the alternating projections algorithm by using the results in [20].

We remark that the sets Ω_1 and Ω_2 given in (10) and (12), respectively, are two C^∞ smooth manifolds that are not closed. The convergence cannot be derived directly by applying the convergence results of the alternating projections between two closed subanalytic sets. By using results in variational analysis and differential geometry, the main convergence results are shown in the following theorem.

Theorem 1 *Let* $\mathbb{M}_i$, $i = 1, .., m$, *and* $\mathbb{M}$ *be the manifolds given in* (8) *and* (6), *respectively. Let* $M \in \mathbb{M}_1 \cap \cdots \cap \mathbb{M}_m \cap \mathbb{M} \neq \emptyset$. *Then there exists a neighborhood* $\mathbb{U}$ *of* M

such that whenever a sequence $\{Z^{(k)}\}$ *derived by Algorithm 1 falls in* $\mathtt{U}$[1]*, and it converges to some* $Z^* \in \mathtt{M}_1 \cap \cdots \cap \mathtt{M}_m \cap \mathtt{M}$ *with rate* $\|Z^{(k)} - Z^*\|_F = O(k^{-\delta})$ *for some* $\delta \in (0, +\infty)$.

To show Theorem 1, it is necessary to study Hölder regularity and separable intersection. For a detailed discussion, we refer to Noll and Rondepierre [20].

Definition 1 [20] Let $\mathtt{A}$ and $\mathtt{B}$ be two sets of points in a Hilbert space equipped with the inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|$. Denote $p_{\mathtt{A}}(x) = \{a \in \mathtt{A} : \|x-a\| = d_{\mathtt{A}}(x)\}$, where $d_{\mathtt{A}}(x) = \min\{\|x-a\| : a \in \mathtt{A}\}$. Similarly, denote $p_{\mathtt{B}}(x) = \{b \in \mathtt{B} : \|x-b\| = d_{\mathtt{B}}(x)\}$, where $d_{\mathtt{B}}(x) = \min\{\|x-b\| : b \in \mathtt{B}\}$. Let $\sigma \in [0, 1)$. The set $\mathtt{B}$ is σ-Hölder regular with respect to $\mathtt{A}$ at $x^* \in \mathtt{A} \cap \mathtt{B}$ if there exists a neighborhood $\mathtt{U}$ of x^* and a constant $c > 0$ such that for every $\bar{y} \in \mathtt{A} \cap \mathtt{U}$ and every $\bar{x} \in p_{\mathtt{B}}(\bar{y}) \cap \mathtt{U}$, one has

$$Ball(\bar{y}, (1+c)r) \cap \{x \mid \bar{y} \in p_{\mathtt{A}}(x), \langle \bar{y} - \bar{x}, x - \bar{x}\rangle > \sqrt{c} r^{\sigma+1} \|x - \bar{x}\|\} \cap \mathtt{B} = \emptyset,$$

where $r = \|\bar{y} - \bar{x}\|$. Note that $p_{\mathtt{B}}(\bar{y})$ is the projection of $\bar{y}$ onto $\mathtt{B}$ and $p_{\mathtt{A}}(x)$ is the projection of x onto $\mathtt{A}$, with respect to the norm. We say that $\mathtt{B}$ is Hölder regular with respect to $\mathtt{A}$ if it is σ-Hölder regular with respect to $\mathtt{A}$ for every $\sigma \in [0, 1)$.

Hölder regularity is mild compared with some other regularity concepts such as prox-regularity [22], Clarke regularity [7] and superregularity [16].

Definition 2 [20] Let $\mathtt{A}$ and $\mathtt{B}$ be two sets of points in a Hilbert space equipped with the inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|$. We say $\mathtt{B}$ intersects separably $\mathtt{A}$ at $x^* \in \mathtt{A} \cap \mathtt{B}$ with exponent $\omega \in [0, 2)$ if there exists a neighborhood $\mathtt{U}$ of x^* such that for every building block $z \to \bar{y} \to \bar{z}$ in $\mathtt{U}$, the condition

$$\langle z - \bar{y}, \bar{z} - \bar{y}\rangle \le (1 - \gamma \|\bar{z} - \bar{y}\|^{\omega}) \|z - \bar{y}\| \|\bar{z} - \bar{y}\| \tag{14}$$

holds with a positive number γ, i.e., it is equivalent to

$$\frac{1 - \cos\alpha}{\|\bar{y} - \bar{z}\|^{\omega}} \ge \gamma,$$

where $\bar{y}$ is a projection point of z onto $\mathtt{A}$, $\bar{z}$ is a projection point of $\bar{y}$ onto $\mathtt{B}$, and α is the angle between $z - \bar{y}$ and $\bar{z} - \bar{y}$.

This separable intersection definition is a new geometric concept that generalizes the transversal intersection [17], the linear regular intersection [15], and the intrinsic transversality intersection [10]. It shows that the definitions of these three kinds of intersections imply $\omega = 0$ in the separable intersection.

The following results are needed to prove our main results.

Theorem 2 (Theorem 1 and Corollary 4 in [20]) *Suppose* $\mathtt{B}$ *intersects* $\mathtt{A}$ *separately at* $x^* \in \mathtt{A} \cap \mathtt{B}$ *with exponent* $\omega \in (0, 2)$*, and* $\mathtt{B}$ *is* $\omega/2$*-Hölder regular at* x^* *with*

[1] $\{Z^{(k)}\}$ falls in $\mathtt{U}$ means there exists a positive integer k_0 such that $\{Z^{(k)}\}_{k \ge k_0} \subset \mathtt{U}$.

respect to $\mathtt{A}$. *Then there exists a neighborhood* $\mathtt{U}$ *of* x^* *such that every sequence of alternating projections between* $\mathtt{A}$ *and* $\mathtt{B}$ *falls in* $\mathtt{U}$, *converges to a point* $x^* \in \mathtt{A} \cap \mathtt{B}$ *with a convergence rate of* $b_k - x^* = O(k^{-\frac{2-\omega}{2\omega}})$ *and* $a_k - x^* = O(k^{-\frac{2-\omega}{2\omega}})$.

Proof of Theorem 1 Let Ω_1 and Ω_2 be given as (10) and (12), respectively. It is clear that finding a point in $\mathtt{M}_1 \cap \cdots \cap \mathtt{M}_m \cap \mathtt{M}$ is equivalent to finding a point X^* in the intersection of Ω_1 and Ω_2. We can set Ω_1 and Ω_2 to be $\mathtt{B}$ and $\mathtt{A}$, respectively, in Theorem 2.

The first task is to show that Ω_1 is Hölder regular with respect to Ω_2 at X^*. Note that Ω_1 is a convex set, and $\pi_{\Omega_1}(Y)$ is single-valued for every $Y \in \mathbb{R}^{n_1 \times \cdots \times n_m} \times \cdots \times \mathbb{R}^{n_1 \times \cdots \times n_m}$. Therefore, Ω_1 is prox-regular. It implies that Ω_1 is $\omega/2$-Hölder regular with respect to Ω_2 at X^* where $\omega \in (0, 2)$.

The next task is to show that Ω_1 intersects separably Ω_2 at $X^* \in \Omega_1 \cap \Omega_2$ with exponent $\omega \in (0, 2)$. Define $f : \Omega_2 \to \mathbb{R}$ as

$$f(X) = \delta_{\Omega_2}(X) + \frac{1}{2} d^2_{\Omega_1}(X), \quad X = (X_1, X_2, ..., X_m) \in \Omega_2, \tag{15}$$

with

$$\delta_{\Omega_2}(X) = \begin{cases} 0 & \text{if } X \in \Omega_2, \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$d_{\Omega_1}(X) = \min\{\|(X - W\|_F : W \in \Omega_1\}.$$

It follows from the definition of $f(X)$ that $f(X^*) = 0$ and X^* is a critical point of f.

Recall that Ω_1 and Ω_2 are two C^∞ manifolds. Then, f is locally Lipschitz continuous, i.e., for each $X \in \Omega_2$, there is an $r > 0$ such that f is Lipschitz continuous on the open ball of center X with radius r. Assume that $(\mathtt{V}, \psi)$ is a local smooth chart of Ω_2 around X^* with bounded $\mathtt{V}$. Therefore, $f(\mathtt{V})$ is bounded by the fact that f is local Lipschitz continuous. According to the definition of the semialgebraic function [18], we can deduce that $f \circ \psi^{-1}$ is also semialgebraic. Then, the Kurdyka-Łojasiweicz inequality [1] for $f \circ \psi^{-1}$ holds for $\bar{W} := \psi(X^*)$. This implies that there exist $\eta \in (0, \infty)$ and a concave function $\tau : [0, \eta]$ such that

(i) $\tau(0) = 0$;
(ii) τ is C^1;
(iii) $\tau' > 0$ on $(0, \eta)$;
(iv) for all $W \in \psi(\mathtt{V}) = \mathtt{U}$ with $f \circ \psi^{-1}(\bar{W}) < f \circ \psi^{-1}(W) < f \circ \psi(\bar{W})) + \eta$, we have

$$\tau'(f \circ \psi^{-1}(W) - f \circ \psi^{-1}(\bar{W}))\, \mathrm{dist}(0, \partial(f \circ \psi^{-1})(W) \geq 1.$$

Moreover, τ is analytic on $\mathtt{V}$; thus, $D(\psi)$ is continuous on $\mathtt{V}$, where D is the differential operator. For every compact subset $\mathtt{K}$ in $\mathtt{V}$, there exists $C_{\mathtt{K}} := \sup_{W \in \mathtt{K}} \|D(\psi(W))\|$, where $\|\cdot\|$ denotes the operator norm. Suppose that $\mathtt{V}'$ is an open set containing X^* in $\mathtt{V}$ such that $\mathtt{K} = cl(\mathtt{V}') \subset int(\mathtt{V})$ is compact ($cl(\mathtt{V}')$ denotes the closure of $\mathtt{V}'$ and $int(\mathtt{V})$

denotes the *interior* of $\mathbb{V}$). Then, for every $X \in \mathbb{V}'$ with $f(X^*) < f(X) < f(X^*)+\eta$, we have

$$C_{\mathrm{K}}\tau'(f(X) - f(X^*))\ \mathrm{dist}(0, \hat{\partial}(f(X)) \geq 1, \tag{16}$$

where $\hat{\partial} f(X)$ is the Fréchet subdifferential of f. We see that the Kurdyka-Łojasiweicz inequality is satisfied for f given in (15).

Here, we construct a function $\tau = t^{1-\theta}$ $(0 < \theta < 1)$ that satisfies (i)-(iv). Because $f(X^*) = 0$, (16) becomes

$$C_{\mathrm{K}}\tau'(f(X))\ \mathrm{dist}(0, \hat{\partial}(f(X)) \geq 1.$$

Since $\tau'(t) = (1-\theta)t^{-\theta}$, there always exists a neighborhood $\mathbb{U}$ of $X^* \in \Omega_1 \cap \Omega_2$ such that $C_{\mathrm{K}}(1-\theta)|f(X)|^{-\theta}\|g\|_F \geq 1$, i.e.,

$$|f(X)|^{-\theta}\|g\|_F \geq c, \quad \text{with } c = \frac{1}{C_{\mathrm{K}}(1-\theta)}, \tag{17}$$

for all $X \in \Omega_2 \cap \mathbb{U}$ and every $g \in \hat{\partial} f(X)$.

By using Algorithm 1, we construct the following sequences according to Definition 2:

$$Z \to \bar{Y} \to \bar{Z} \to \cdots.$$

Here, $\bar{Y}$ is the projection $\pi_{\Omega_2}(Z)$ and $\bar{Z}$ is the projection $\pi_{\Omega_1}(\bar{Y})$, with $\pi_{\Omega_1}(\cdot)$ and $\pi_{\Omega_2}(\cdot)$ being defined as (11) and (13), respectively. Suppose Z and $\bar{Z}$ are in $\mathbb{U}$, $\bar{Y} \in \mathbb{U} \cap \Omega_2$; we obtain the proximal normal cone to Ω_2 at $\bar{Y}$:

$$\mathbb{N}^p_{\Omega_2}(\bar{Y}) = \{\lambda V : \lambda \geq 0,\ \bar{Y} \in \pi_{\Omega_2}(\bar{Y} + V)\}.$$

According to the definition of the Fréchet subdifferential, $G \in \hat{\partial} f(\bar{Y})$ if and only if $G = V + \bar{Y} - \bar{Z}$ for every $V \in \mathbb{N}^p_{\Omega_2}(\bar{Y})$ of the form $V = \lambda(Z - \bar{Y})$.

Note that $\bar{Y} \in \pi_{\Omega_2}(Z)$, from (15), we have $f(\bar{Y}) = \frac{1}{2}d^2_{\Omega_1}(\bar{Y})$. Substitute $f(\bar{Y})$ into (17) to obtain

$$2^{\theta} d_{\Omega_1}(\bar{Y})^{-2\theta}\|\lambda(Z - \bar{Y}) + (\bar{Y} - \bar{Z})\|_F \geq c > 0,$$

for every $\lambda \geq 0$. It follows that

$$d_{\Omega_1}(\bar{Y})^{-2\theta} \min_{\lambda \geq 0} \|\lambda(Z - \bar{Y}) + (\bar{Y} - \bar{Z})\|_F \geq 2^{-\theta} c. \tag{18}$$

Let α be the angle between the iterates, which can be defined as the angle between $Z - \bar{Y}$ and $\bar{Z} - \bar{Y}$. Let us consider two cases:

(i) When $0 < \alpha \leq \pi/2$, we have

$$\min_{\lambda \geq 0} \|\lambda(Z - \bar{Y}) + (\bar{Y} - \bar{Z})\|_F = \|\bar{Y} - \bar{Z}\|_F \sin\alpha.$$

By substituting it into (18), we obtain

$$\frac{\sin\alpha}{d_{\Omega_1}(\bar{Y})^{2\theta-1}} \geq 2^{-\theta}c.$$

Note that $1-\cos\alpha \geq \frac{1}{2}\sin^2\alpha$. We have

$$\frac{1-\cos\alpha}{d_{\Omega_1}(\bar{Y})^{4\theta-2}} \geq 2^{-2\theta-1}c^2. \tag{19}$$

When the numerator tends to 0, the denominator has to go to zero, which implies that $4\theta-2>0$, i.e., $\theta>\frac{1}{2}$.

(ii) When $\pi/2<\alpha<\pi$, we have $\cos\alpha<0$, i.e., $1-\cos\alpha\geq 1$. The infimum in (18) is attained at $\lambda=0$, and (18) becomes $d_{\Omega_1}(\bar{Y})^{1-2\theta}\geq 2^{-\theta}c$. Therefore,

$$\frac{1}{d_{\Omega_1}(\bar{Y})^{4\theta-2}} \geq 2^{-2\theta}c^2 > 2^{-2\theta-1}c^2.$$

Since $1-\cos\alpha\geq 1$, we have $\frac{1-\cos\alpha}{d_{\Omega_1}(\bar{Y})^{4\theta-2}}\geq 2^{-2\theta-1}c^2$, i.e., (19) is satisfied with $\theta>\frac{1}{2}$.

Therefore, Ω_1 intersects Ω_2 separably with the exponent $\omega=4\theta-2\in(0,2)$, the corresponding number γ in Definition 2 can be set to be $2^{-2\theta-1}c^2$. By Theorem 2, the result follows by setting $\delta=(2-\omega)/2\omega\in(0,+\infty)$. □

In the next section, we test our method and nonnegative tensor decomposition methods on synthetic data and real-world data. The results show that the performance of the proposed alternating projections method is better than the others.

4 Experimental results

4.1 Methods compared

We compare the following state-of-the-art Nonnegative Tucker decomposition (NTD) methods for the nonnegative tensor decomposition:

NTD-HALS: An HALS algorithm [32]
NTD-MU: A multiple updating algorithm [32]
NTD-BCD: A block coordinate descent method [30]
NTD-APG: An accelerated proximal gradient algorithm [32]

We also compare the proposed model with a well-known nonnegative CANDECOMP/PARAFAC decomposition (NCPD), that is, given a tensor $\mathcal{A} \in \mathbb{R}_+^{n_1 \times n_2 \times \cdots \times n_m}$,

$$\begin{aligned} &\min \|\mathcal{A} - \sum_{z=1}^{Z} \lambda_z \mathbf{a}^{z,1} \otimes \mathbf{a}^{z,2} \otimes \cdots \mathbf{a}^{z,m}\|_F^2, \\ &\text{s.t.} \quad \boldsymbol{\lambda} = \left(\lambda_1 \cdots \lambda_Z\right) \geq 0, \quad \mathbf{A}^t = \left(\mathbf{a}^{1,t} \cdots \mathbf{a}^{Z,t}\right) \geq 0 \quad (t = 1, \ldots, m). \end{aligned} \tag{20}$$

The state-of-the-art methods for the NCPD model are presented as follows.

NCPD-HALS: A hierarchical ALS algorithm [5, 6]
NCPD-MU: A fixed point (FP) algorithm with multiplicative updating [28]
NCPD-BCD: A block coordinate descent (BCD) method [30]
NCPD-APG: An accelerated proximal gradient method [31]
NCPD-CDTF: A block coordinate descent method [23]
NCPD-SaCD: A saturating coordinate descent method with the Lipschitz continuity-based element importance updating rule [2]

We list the computational cost of these methods in Table 1. The cost of the proposed NLRT method per iteration is approximately the same as that of the NTD-type methods. As they involve the calculation of nonnegative vectors only, the cost of the NCP-type methods per iteration is smaller than that of the proposed NLRT method.

The stopping criterion of the proposed method and the other comparison methods is that the relative difference between successive iterates is smaller than 10^{-5}. All the experiments are conducted on an Intel(R) Core(TM) i9-9900K CPU@3.60 GHz with 32 GB of RAM using MATLAB. Throughout this section, we mainly test the low-rank approximation ability of our method and the nonnegative tensor decomposition methods with a given rank. That is, the CP rank and the multilinear rank are manually prescribed. For real-world applications, we suggest two adaptive rank adjusting strategies proposed in [29]. The basic idea is to use a large (or a small) value of the rank as the initial guess and adaptively decrease (or increase) the rank based on the QR decomposition of the unfolding matrices as the algorithm iterates. The effectiveness of those strategies has been demonstrated in [29].

4.2 Synthetic data sets

We first test different methods on synthetic data sets. We generate two kinds of synthetic data as follows:

– Case 1 (Noisy nonnegative low-rank tensor): We generate low rank nonnegative tensors in two steps. First, a core tensor of size $r_1 \times r_2 \times \ldots \times r_m$ (i.e., the multilinear rank is $(r_1, r_2, \cdots, r_m)$) and m factor matrices of sizes $n_i \times r_i$ $(i = 1, 2, \cdots, m)$ are generated with the entries uniformly distributed in [0, 1]. Second, these factor matrices are multiplied by the core tensor via the tensor-matrix product to generate the low rank nonnegative tensors of size $n_1 \times n_2 \times \cdots \times n_m$, and each entry is elementwisely divided by the maximal value, being in the interval of [0, 1]. Finally,

Table 1 The computational cost

Method	Complexity	Details of the most expensive computations
NCPD-MU	$O(mr\prod_{j=1}^{m} n_j)$	Khatri-Rao product and unfolding matrices times Khatri-Rao product
NCPD-HALS	$O(mr\prod_{j=1}^{m} n_j)$	Khatri-Rao product and unfolding matrices times Khatri-Rao product
NCPD-BCD	$O(mr\prod_{j=1}^{m} n_j)$	Khatri-Rao product and unfolding matrices times Khatri-Rao product.
NCPD-APG	$O(mr\prod_{j=1}^{m} n_j)$	Khatri-Rao product and unfolding matrices times Khatri-Rao product
NCPD-CDTF	$O(m^2 r\prod_{j=1}^{m} n_j)$	Khatri-Rao product of rank one components and vectors times Khatri-Rao product.
NCPD-SaCD	$O(mr\prod_{j=1}^{m} n_j)$	Khatri-Rao product and unfolding matrices times Khatri-Rao product
NTD-MU	$O(\sum_{i=1}^{m}\prod_{j\neq i}^{m} n_j r_i^2)$	MU on unfolding matrices $\{\mathbf{A}_i\}_{k=1}^{m}$.
NTD-HALS	$O(\sum_{i=1}^{m}\prod_{j\neq i}^{m} n_j r_i)$	HALS on unfolding matrices $\{\mathbf{A}_i\}_{k=1}^{m}$
NTD-BCD	$O(\sum_{i=1}^{m}\prod_{j\neq i}^{m} n_j r_i(r_i+n_i))$	The tensor-matrix multiplication and the matrix multiplication between the i-th unfolding matrix of $\mathcal{G}\times_{j=1,j\neq i}\mathbf{U}^{(j)}$ and its transpose
NTD-APG	$O(\sum_{i=1}^{m}\prod_{j\neq i}^{m} n_j r_i^2)$	The tensor-matrix multiplications among a) the i-th factor matrix b) the transpose of the i-th unfolding matrix of $\mathcal{G}\times_{j=1,j\neq i}\mathbf{U}^{(j)}$ and c) the i-th unfolding matrix of $\mathcal{G}\times_{j=1,j\neq i}\mathbf{U}^{(j)}$
NLRT	$O((\prod_{j=1}^{m} n_j)\sum_{i=1}^{m} r_i)$	SVDs of unfolding matrices $\{\mathbf{A}_i\}_{k=1}^{m}$

we add Gaussian noise to generate noisy tensors with different signal-to-noise ratio (SNR)[2].

– Case 2 (Nonnegative random tensor): We randomly generate nonnegative tensors of size $n_1\times n_2\times\cdots\times n_m$, where their entries follow a uniform distribution between 0 and 1. The tensor data are fixed once generated, and the low rank minimizer is unknown in this setting. For CP decomposition methods, the CP rank is set to r. For Tucker decomposition methods, the multilinear rank is set to be $[r, r, \ldots, r]$.

It is not straightforward to make the comparison between the NCPD methods with low multilinear rank-based methods fair, owing to different definitions of the rank. For NCPD methods, determining the CP rank of a given tensor is NP-hard [13]. Fortunately, we have that given the multilinear rank $(r_1, r_2, \ldots, r_m)$ of a tensor, its CP rank cannot be larger than $\prod_{k=1}^{m} r_k$. Therefore, in Case 1, we select the CP rank in the NCPD methods from a set with three candidates, i.e., $\{\prod_{k=1}^{m} r_k, \sum_{k=1}^{m} r_k, \max_i r_i\}$. Then, we report the best relative approximation error in the NCPD methods. We believe this makes the comparison with the NCPD methods possible and fair to a certain extent in

[2] To avoid making the entries negative, we first simulate noise with a standard normal distribution and then set the negative noisy value to 0. The SNR in dB is defined as $\mathrm{SNR}_{\mathrm{dB}} = 20\log_{10}\frac{\|\mathcal{X}_{\mathrm{groundtruth}}\|_F}{\|\mathrm{Noise}\|_F}$.

Case 1. In Case 2, we set the CP rank as r for the NCPD methods when the multilinear rank is $[r, r, \ldots, r]$. In this situation, the results by the NCPD methods only reflect the representation ability of these NCPD methods.

We report the relative approximation error[3] to quantitatively measure the approximation quality. The ground truth tensor is the generated tensor without noise. The relative approximation errors of the results by different methods in Case 1 are reported in Table 2. The reported entries of all the comparison methods in the table are the average values together with the standard deviations of ten trials with different random initial guesses in the CP decomposition vectors and the Tucker decomposition matrices. However, the results of the proposed NLRT method are deterministic when the input nonnegative tensor is fixed. We can see from Table 2 that the proposed NLRT method achieves the best performance and is also quite robust to different noise levels.

In Table 3, we report the average running time of each method. For tensors with the same size, the NCPD methods and NTD methods need the same computation time for different noise levels. The running time of our NLRT becomes less when the SNR value is larger. This indicates that our method could converge faster with less noise. Meanwhile, we can see that as the number of total elements in the tensor grows from 10^6 ($100 \times 100 \times 100$) to 2.3×10^7 ($30 \times 30 \times 30 \times 30 \times 30$), the running time of all the methods increases rapidly. Since our method involves SVD computations, whose computational complexity grows cubically with the dimension, our superior efficiency is obvious for smaller data.

The relative approximation errors in Case 2 with respect to different values of r are plotted in Fig. 1. As we stated, the tensor of a given size will be fixed once generated. Then, for different values of r, we run each algorithm 10 times, and the averaged values are plotted. From Fig. 1, we can see that the proposed NLRT method and NTD-BCD perform better than the other methods. For tensors of size $40 \times 40 \times 40$, the superiority of our method over NTD-BCD is obvious when the rank is between 27 and 39.

4.3 Video data

In this subsection, we select 5 videos[4] to test our method on the task of approximation. Three videos (named "foreman", "coastguard" and "news") are of size $144 \times 176 \times 100$ (height$\times$width$\times$frame), and one (named "basketball") is of size $44 \times 256 \times 40$. One long video (named "bridge-far") of size $144 \times 176 \times 2000$ is also selected to test the approximation ability for large-scale data. First, we set the multilinear rank to be $(r, r, \cdots, r)$ and the CP rank to be r. We test our method to approximate these five videos with varying r from 5 to 100. Moreover, we add Gaussian noise to the video "coastguard" with different noise levels ($\text{SNR}_{\text{dB}} = 20, 30, 40, 50$) and test the approximation ability of different methods for noisy video data.

We plot the relative approximation errors with respect to r on 5 videos in Figs. 2 and 3. Although, for some videos the approximation errors of the results by the NCPD methods are much higher than those for others, because setting CP rank to r

[3] Defined as $\frac{\|\mathcal{X}_{\text{estimated}} - \mathcal{X}_{\text{groundtruth}}\|_F}{\|\mathcal{X}_{\text{groundtruth}}\|_F}$.

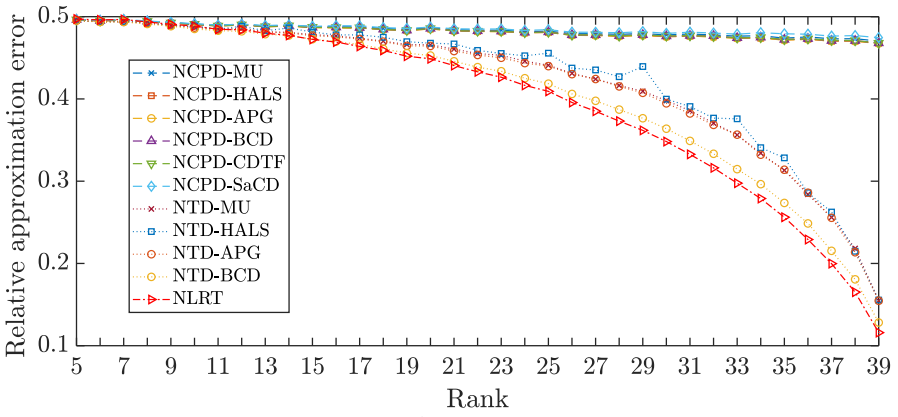
[4] Videos are available at http://trace.eas.asu.edu/yuv/ and https://sites.google.com/site/jamiezeminzhang/publications.

Table 2 The mean values (and standard deviations) of the relative approximation errors of the results by different methods in Case 1. The **best** values are highlighted in bold. (The mean values and standard deviations are shown in percentages)

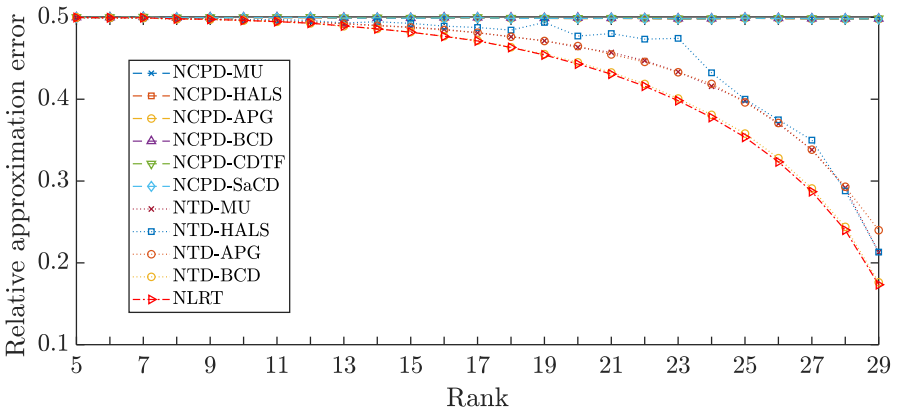
Tensor size: $100 \times 100 \times 100$								Multilinear rank: [5, 5, 5]				
SNR (dB)	Noisy	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
30	3.16	2.86	2.78	2.74	2.74	2.75	2.95	2.84	2.75	2.73	2.75	**2.73**
		(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.11)	(0.05)	(0.03)	(0.00)	(0.01)	
40	1.00	1.21	1.01	0.87	0.87	0.87	1.31	1.11	1.00	0.88	0.95	**0.86**
		(0.02)	(0.01)	(0.00)	(0.00)	(0.00)	(0.14)	(0.15)	(0.12)	(0.01)	(0.05)	
50	0.32	0.91	0.59	0.28	0.28	0.28	0.97	0.67	0.50	0.33	0.51	**0.27**
		(0.02)	(0.03)	(0.00)	(0.00)	(0.00)	(0.21)	(0.19)	(0.16)	(0.01)	(0.07)	
30	3.16	2.94	2.77	2.74	2.75	2.75	2.99	2.68	2.67	2.67	2.67	**2.66**
		(0.01)	(0.00)	(0.00)	(0.01)	(0.01)	(0.16)	(0.01)	(0.00)	(0.00)	(0.00)	
40	1.00	1.40	0.96	0.87	0.88	0.88	1.41	0.91	0.88	0.86	0.85	**0.84**
		(0.03)	(0.02)	(0.00)	(0.02)	(0.01)	(0.13)	(0.04)	(0.02)	(0.01)	(0.01)	
50	0.32	1.14	0.52	0.29	0.34	0.32	1.22	0.41	0.35	0.31	0.31	**0.27**
		(0.03)	(0.03)	(0.01)	(0.09)	(0.04)	(0.29)	(0.05)	(0.03)	(0.02)	(0.03)	
Tensor size: $30 \times 30 \times 30 \times 30 \times 30$								Multilinear rank: [2, 2, 2, 2, 2]				
SNR (dB)	Noisy	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
30	3.16	2.98	2.77	2.74	2.76	2.77	3.08	2.48	2.48	2.47	2.48	**2.48**
		(0.07)	(0.01)	(0.00)	(0.01)	(0.01)	(0.17)	(0.00)	(0.01)	(0.00)	(0.00)	
40	1.00	1.11	0.89	0.87	0.90	0.89	1.63	0.83	0.81	0.81	0.81	**0.80**
		(0.06)	(0.02)	(0.00)	(0.03)	(0.02)	(0.33)	(0.07)	(0.01)	(0.01)	(0.01)	
50	0.32	0.75	0.38	0.28	0.35	0.39	1.19	0.28	0.28	0.27	0.28	**0.25**
		(0.09)	(0.03)	(0.01)	(0.05)	(0.09)	(0.38)	(0.02)	(0.01)	(0.01)	(0.03)	

Table 3 The averaged running time (in seconds) of different methods in Case 1

Tensor size: $100 \times 100 \times 100$						Multilinear rank: [5, 5, 5]					
SNR (dB)	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
30	6.6	0.5	5.0	2.2	1.1	6.1	12.2	12.7	16.6	5.4	0.5
40	6.4	0.5	5.0	13.0	14.7	6.3	12.1	12.8	16.7	5.4	0.4
50	6.6	0.5	9.2	13.0	15.3	6.8	12.2	12.9	16.6	5.5	0.3
Tensor size: $50 \times 50 \times 50 \times 50$						Multilinear rank: [3, 3, 3, 3]					
SNR (dB)	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
30	61.5	40.3	108.2	112.0	139.7	36.2	16.2	16.0	23.1	33.7	11.9
40	60.2	39.0	106.9	112.3	137.6	35.8	16.3	15.9	23.1	41.1	8.3
50	60.2	47.9	106.3	103.0	147.0	36.2	16.0	16.1	22.7	40.9	5.8
Tensor size: $30 \times 30 \times 30 \times 30 \times 30$						Multilinear rank: [2, 2, 2, 2, 2]					
SNR (dB)	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
30	249.4	159.7	215.4	218.7	195.3	127.2	115.0	119.6	120.4	102.6	106.2
40	219.5	192.7	150.7	215.3	224.4	130.4	112.6	117.3	118.9	123.3	78.6
50	233.4	184.1	127.9	243.4	324.2	129.3	114.9	119.5	121.1	131.0	56.1



(a) Tensor size: $40 \times 40 \times 40$



(b) Tensor size: $30 \times 30 \times 30 \times 30$

Fig. 1 Relative approximation errors on the randomly generated tensors in Case 2 with respect to the different rank settings

largely constrain s the model's representation ability; however, we can still see that the potential of the NCPD methods are promising. For example, for the videos "news" and "bridge-far", the NCPD methods are even occasionally superior to the NTD methods. Thus, the comparison with the NCPD methods provides some insights. From Figs. 2 and 3, it can be seen that the approximation errors of the results by our method are the lowest. Fig. 4 shows the relative approximation errors on the noisy video "coastguard" with respect to r. Similarly, our method achieves the lowest approximation errors on the video "coastguard" with respect to different rank settings and different noise levels. In Table 4, we list the average running time of each method.

4.4 Hyperspectral data

In this subsection, we test the different methods on hyperspectral data. We consider four hyperspectral images (HSIs): a subimage of the Pavia City Center[5] data set of size

[5] Data available at http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

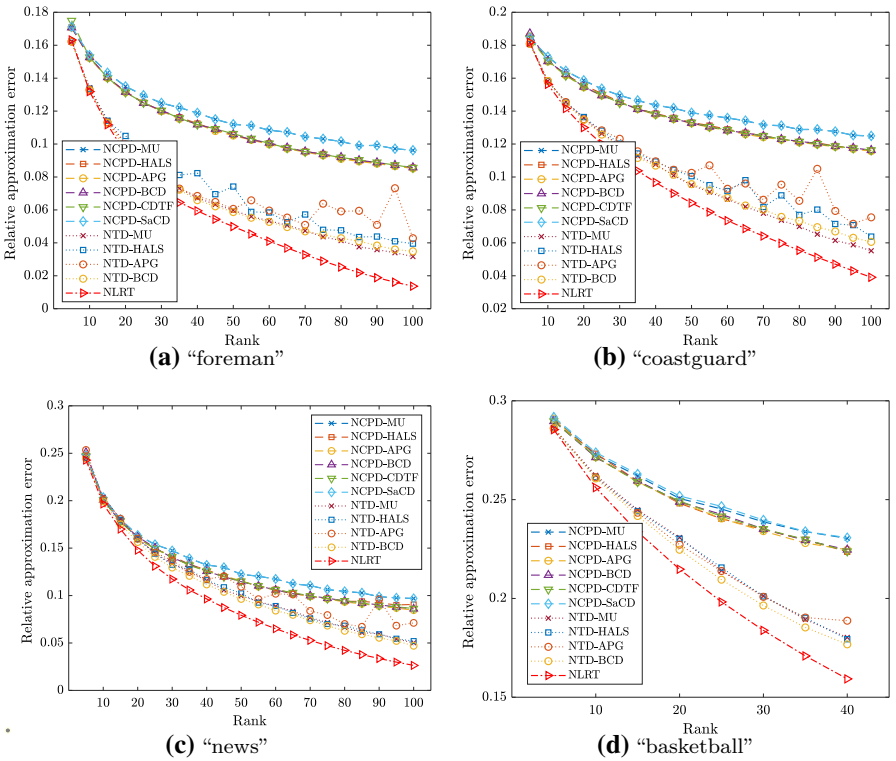


Fig. 2 Relative approximation errors on 4 videos (100 frames) with respect to the different rank settings

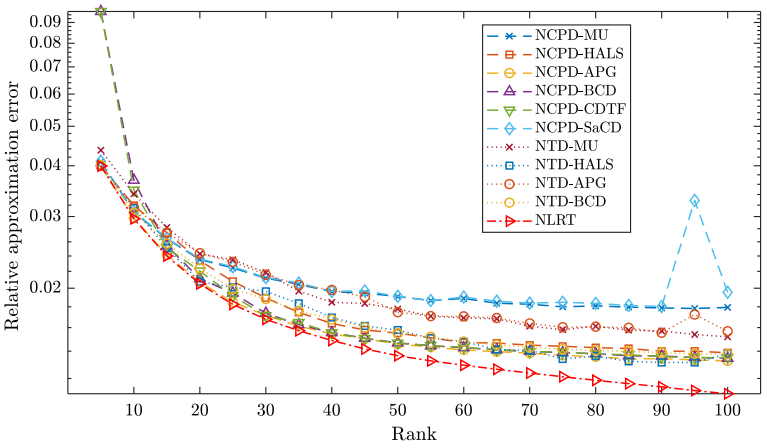


Fig. 3 Relative approximation errors on the video "bridge-far" (2000 frames) with respect to the different rank settings

Table 4 The average running time (in seconds) of the different methods on the video data

Video	# frames	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
"foreman"	100	60	66	55	24	23	45	202	188	354	74	20
"news"	100	46	46	37	17	16	33	176	197	313	59	25
"coastguard"	100	35	36	30	13	12	27	129	165	228	46	15
"basketball"	40	16	17	11	3	3	12	25	20	34	14	15
"bridge-far"	2000	386	173	265	186	188	209	183	211	299	511	296
Video	SNR (dB)	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
"coastguard"	20	28	29	28	9	9	20	135	146	258	46	18
	30	28	29	29	10	10	20	133	145	254	46	17
	40	28	29	29	10	10	20	134	147	255	47	16
	50	28	29	29	10	9	20	136	143	259	46	15

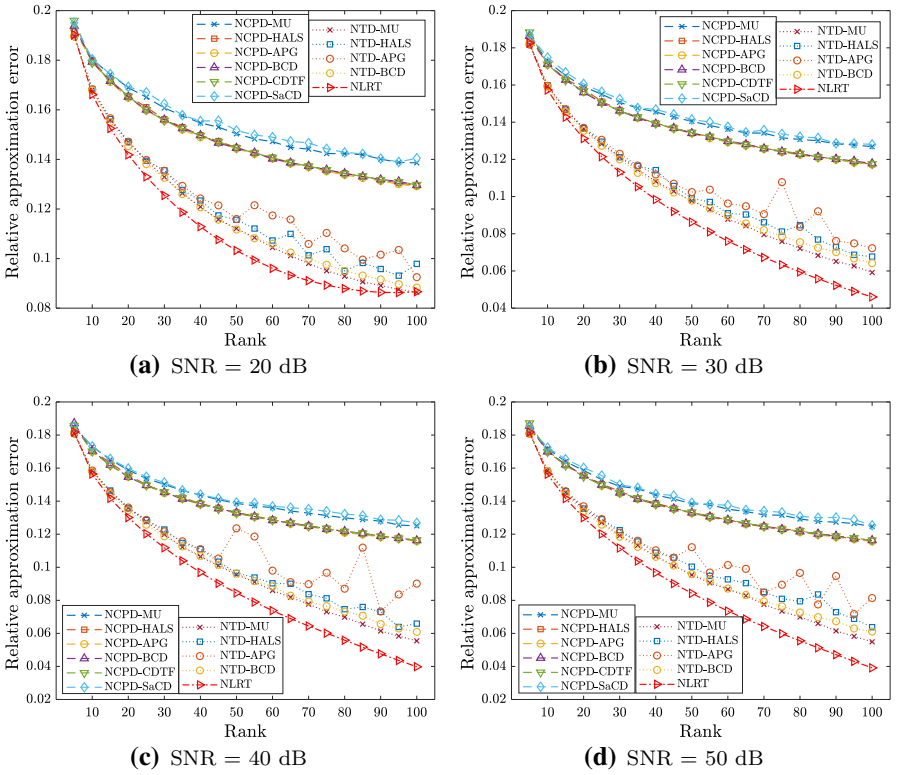


Fig. 4 Relative approximation errors on the noisy video "coastguard" with respect to different rank settings and different noise levels

$200 \times 200 \times 80$ (height$\times$width$\times$spectrum), a subimage of the Washington DC Mall[6] data set of size $200 \times 200 \times 160$, the RemoteImage[7] of size $200 \times 200 \times 89$, and a subimage of the Curprite[8] data set of size $150 \times 150 \times 150$. Meanwhile, a hyperspectral video (HSV)[9] of size $120 \times 188 \times 33 \times 31$ (height$\times$width$\times$spectrum$\times$time) is also selected to test the effectiveness of the different methods on a fourth-order tensor.

Figs. 5 and 6 report the relative approximation errors with respect to different values of rank r, i.e., multilinear rank $= (r, r, r)$ or (r, r, r, r) and CP rank $= r$. It is evident that the relative approximation errors by our NLRT are the lowest among all the methods. It is interesting to note that the difference between our method and NTD-BCD (the second best method in our comparison) is more significant than that on the synthetic fourth-order tensor data.

In Fig. 7, we display the pseudocolor images of the results for the Washington DC Mall data set with a multilinear rank (100,100,100) and a CP rank 100. The pseudocolor image is composed of the 113-th, 2-nd, and 16-th bands as the red, green

[6] Data available at https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html.

[7] Data available at https://www.cs.rochester.edu/~jliu/code/TensorCompletion.zip.

[8] Data available at https://aviris.jpl.nasa.gov/data/free_data.html.

[9] Data available at http://openremotesensing.net/knowledgebase/hyperspectral-video/.

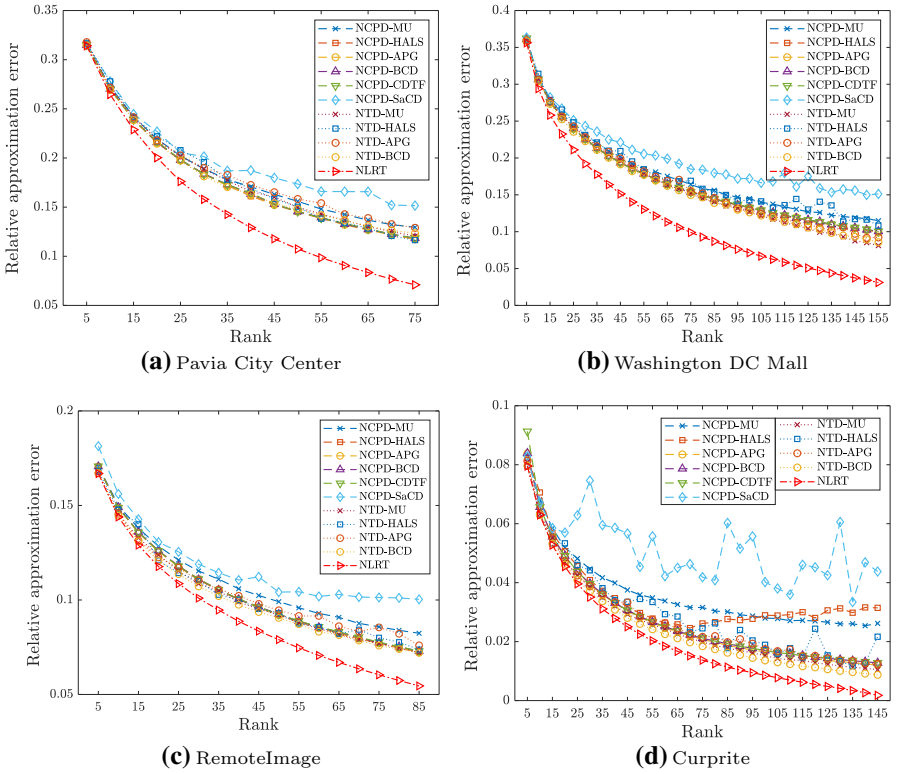


Fig. 5 Relative approximation errors on 4 HSIs with respect to the different rank settings

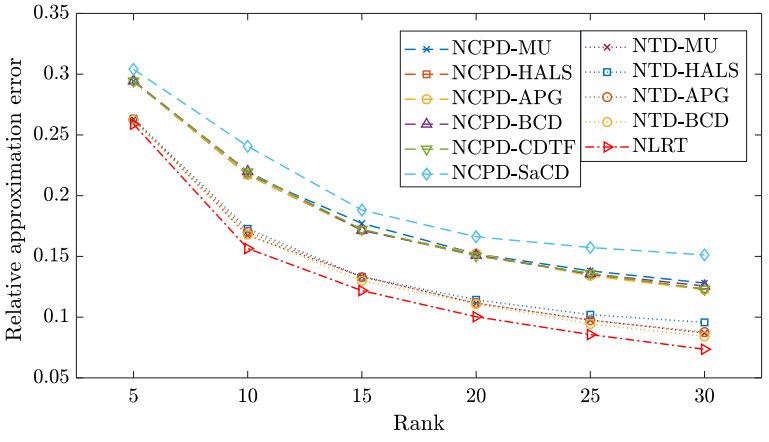
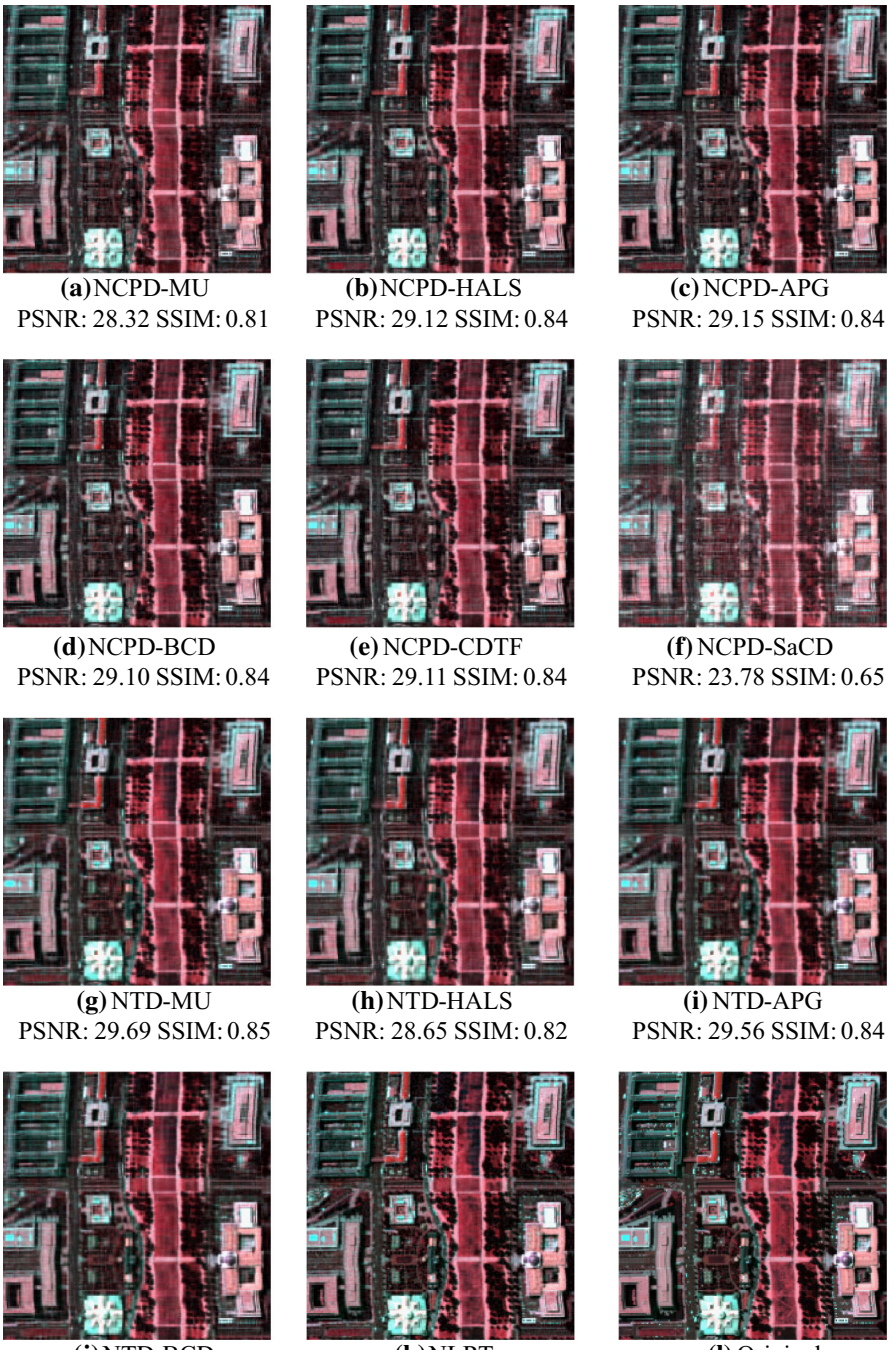


Fig. 6 Relative approximation errors on the HSV with respect to the different rank settings



(a) NCPD-MU
PSNR: 28.32 SSIM: 0.81

(b) NCPD-HALS
PSNR: 29.12 SSIM: 0.84

(c) NCPD-APG
PSNR: 29.15 SSIM: 0.84

(d) NCPD-BCD
PSNR: 29.10 SSIM: 0.84

(e) NCPD-CDTF
PSNR: 29.11 SSIM: 0.84

(f) NCPD-SaCD
PSNR: 23.78 SSIM: 0.65

(g) NTD-MU
PSNR: 29.69 SSIM: 0.85

(h) NTD-HALS
PSNR: 28.65 SSIM: 0.82

(i) NTD-APG
PSNR: 29.56 SSIM: 0.84

(j) NTD-BCD
PSNR: 29.78 SSIM: 0.85

(k) NLRT
PSNR: **34.56** SSIM: **0.94**

(l) Original
PSNR: 28.32 SSIM:0.81

Fig. 7 Pseudocolor images composed of the 113-th, 2-nd, and 16-th bands of the nonnegative low-rank approximations by the different methods when setting the rank 100 on the Washington DC Mall

and blue channels, respectively. We also compute two image quality assessments (IQAs): the peak signal-to-noise ratio (PSNR)[10] and the structural similarity index (SSIM) [27] of all the spectral bands for each band. Higher values of these two indices indicate a better reconstruction quality. In Fig. 7, we report the mean values across the spectral bands of these two IQAs. It can be found in Fig. 7 that both the visual and quality assessments of the NCPD methods are comparable to the NTD methods. The proposed NLRT method largely outperforms the other methods in terms of two IQAs, achieving first place.

4.5 Selection of features

One advantage of the proposed NLRT method is that it can provide a significant index based on the singular values of the unfolding matrices [25] that can be used to identify important singular basis vectors in the approximation. Those singular values and singular vectors are natural concomitants brought out by our algorithm without additional computations of the SVD.

Here, we take the HSI Washington DC Mall as an example. We compute the low-rank approximations of the proposed NLRT method and the other methods in our comparison with multilinear rank (r, r, r) and CP rank r for $r = 20, 40, 60, 80, 160$. For the approximation results by NCPD methods, we normalize the base vectors in (20) such that the ℓ_2 norms of $\mathbf{a}^{k,1}$, $\mathbf{a}^{k,2}$ and $\mathbf{a}^{k,3}$ are equal to 1, and rearrange the resulting values λ'_z in descending order in the CP decomposition. In Fig. 8, we plot

$$\|\mathcal{A} - \mathcal{X}_{\mathrm{NCPD}}(j)\|_F / \|\mathcal{A}\|_F$$

with respect to j, where $\mathcal{X}_{\mathrm{NCPD}}(j) = \sum_{k=1}^{j} \lambda'_k \mathbf{a}^{k,1} \otimes \mathbf{a}^{k,2} \otimes \mathbf{a}^{k,3}$. Similarly, for the results of the NTD methods, we also plot

$$\|\mathcal{A} - \mathcal{X}_{\mathrm{NTD}}(j)\|_F / \|\mathcal{A}\|_F$$

with respect to j, where $\mathcal{X}_{\mathrm{NTD}}(j) = [\mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}]_{:,:,\mathbf{k}_j} \times_3 \mathbf{U}^{(3)}_{:,\mathbf{k}_j}$, $[\mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}]_{:,:,\mathbf{k}_j}$ is the $\mathbf{k}_j$-th mode-12 (spatial) slice of $[\mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}]$, and each $[\mathcal{G} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}]_{:,:,\mathbf{k}_j}$ is normalized with its Frobenius norm equal to 1, and $\mathbf{k}_j$ indicates a vector composed of the indices corresponding to the j largest ℓ_2 norms of $\mathbf{U}^{(3)}$'s columns. For the results by our methods, we plot

$$\|\mathcal{A} - \mathcal{X}_{\mathrm{NLRT}}(j)\|_F / \|\mathcal{A}\|_F$$

with respect to j, where $\mathcal{X}_{\mathrm{NLRT}}(j) = \mathrm{fold}\left(\sum_{i=1}^{j} \sigma_i(\mathbf{X}_3)\mathbf{u}_i(\mathbf{X}_3)\mathbf{v}_i^T(\mathbf{X}_3)\right)$, $\sigma_i(\mathbf{X}_3)$ is the i-th singular value of $\mathbf{X}_3$, and $\mathbf{X}_3$ is the third-mode unfolding matrix of $\mathcal{X}$. The third mode of X is chosen in NTD and our NLRT, and we are interested in observing how many indices are required in the spectral mode of the given hyperspectral data.

[10] https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

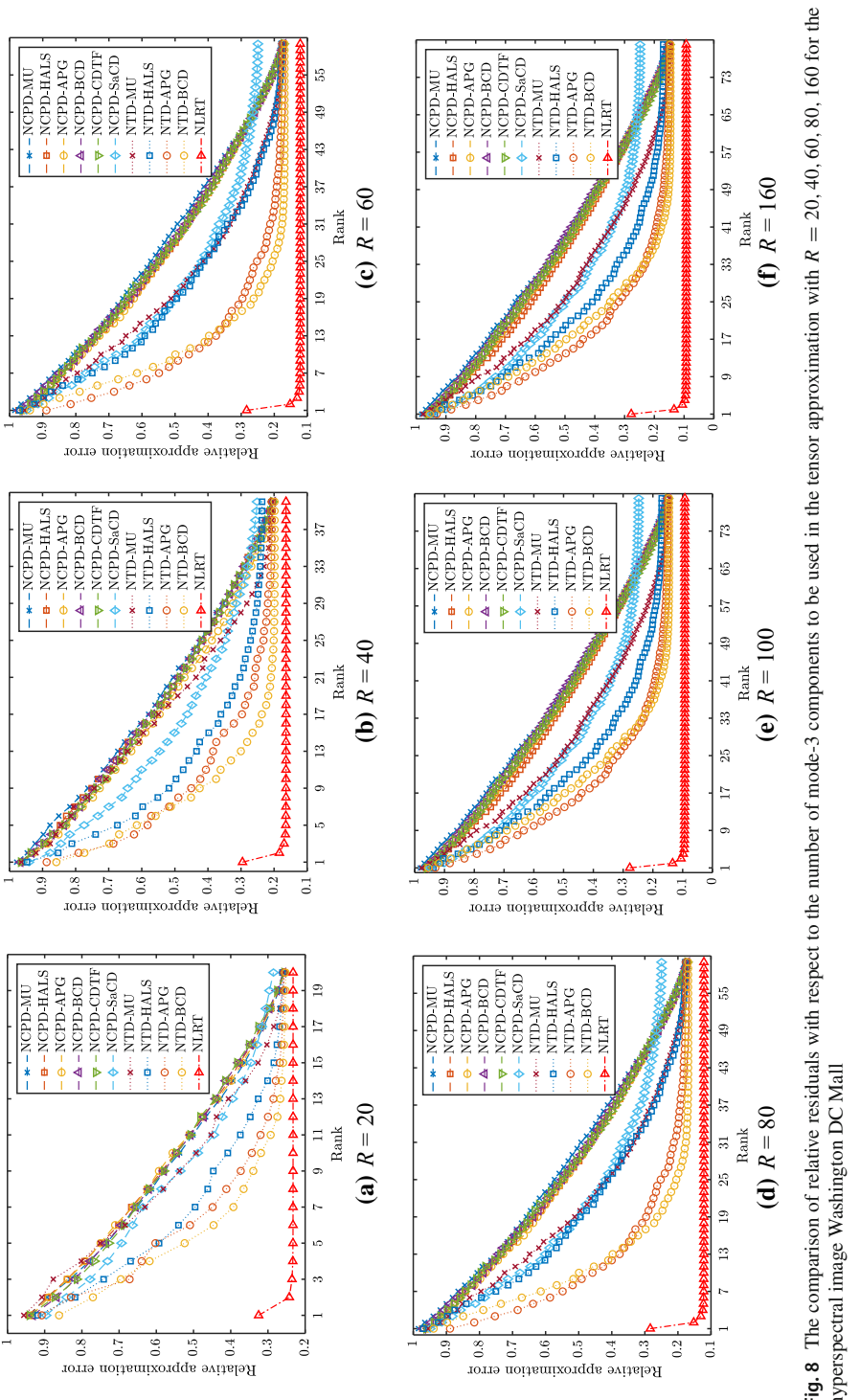


Fig. 8 The comparison of relative residuals with respect to the number of mode-3 components to be used in the tensor approximation with $R = 20, 40, 60, 80, 160$ for the hyperspectral image Washington DC Mall

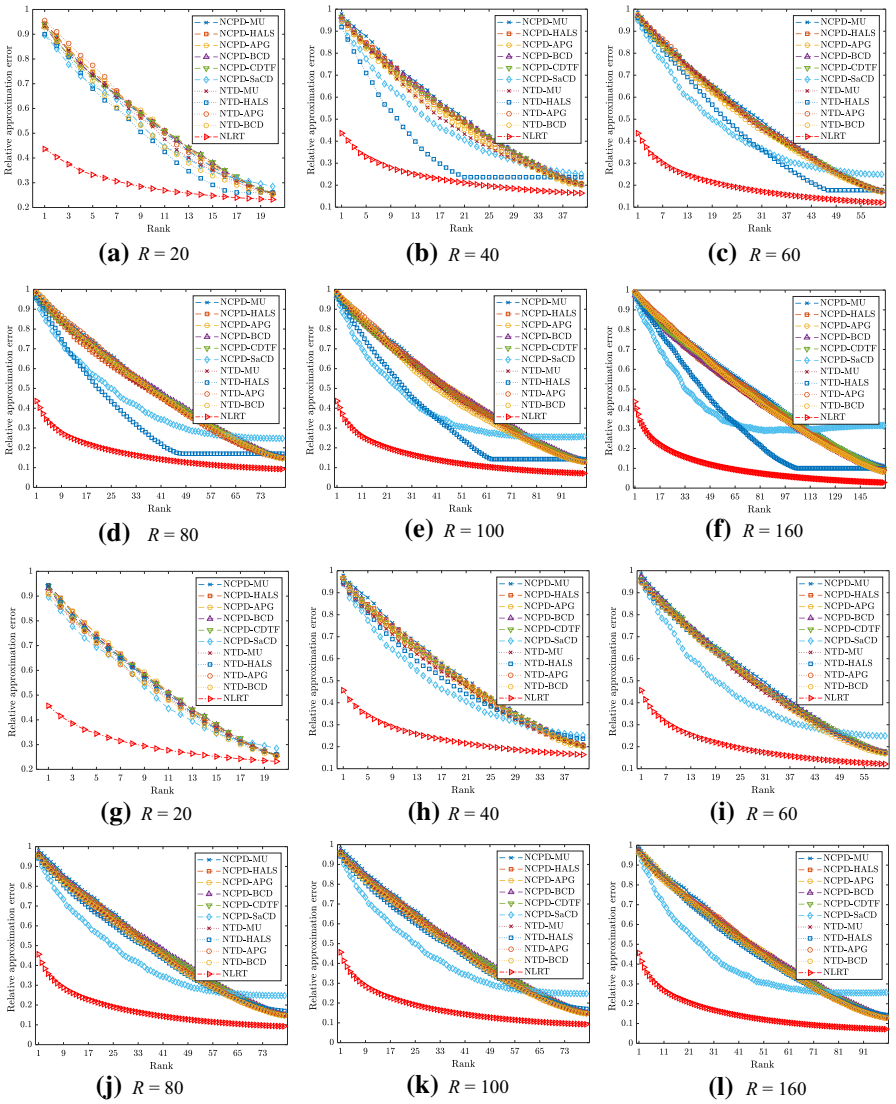
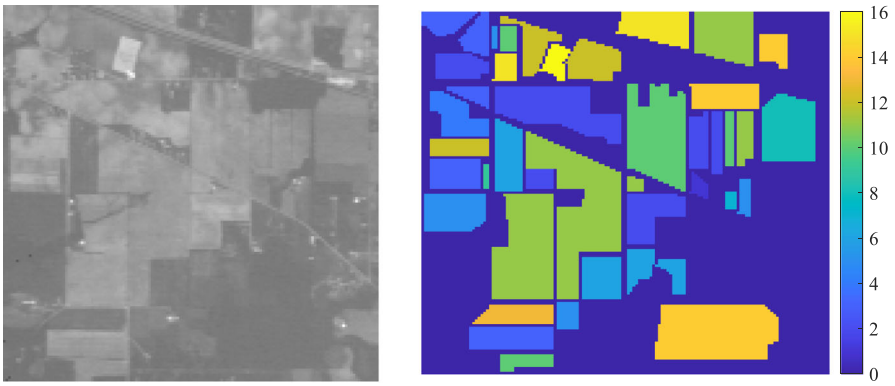


Fig. 9 The comparison of relative residuals with respect to the number of the first mode (upper two rows from (a) to (f)) and the second mode (bottom two rows from (g) to (l)) components to be used in the tensor approximation with $R = 20, 40, 60, 80, 160$ for the hyperspectral image Washington DC Mall

In Fig. 8, we can see that when the number of components (namely, j) increases, the relative residual decreases. Our NLRT could provide a significant index based on singular values to identify important singular basis vectors for the approximation. Thus, the relative residuals by the proposed NLRT algorithm are significantly smaller than those of the NTD and NCPD algorithms. Similar phenomena can be found in Fig. 9, in which $\mathcal{X}_{\text{NTD}}(j)$ and $\mathcal{X}_{\text{MP-NLRT}}(j)$ are computed using the number of indices in the first or second modes of X.



(a) The 10-th band of the original HSI. **(b)** The ground truth categorization map.

Fig. 10 Indian Pines image and the related ground truth categorization information

4.6 Image classification

The advantage of the proposed NLRT method is that the important singular basis vectors can be identified within the algorithm. Such basis vectors can provide useful information for image recognitions such as classification. Here, we conduct hyperspectral image classification experiments on the Indian Pines[11] data set. This data set was captured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in northwestern Indiana in June 1992. After removing 20 bands, which cover the region of water absorption, this HSI is of size $145 \times 145 \times 200$. The ground truth contains 16 land cover classes, as shown in Fig. 10. Therefore, we set the multilinear rank to (16, 16, 16) and the CP rank to 16 for all the methods compared. We randomly choose s of the available labeled samples, which are exhibited in Table 5. Labeled samples from each class are used for training, and the remaining samples are used for testing.

After obtaining low rank approximations, 16 singular vectors corresponding to the largest 16 singular values of the unfolding matrix of the tensor approximation along the spectral mode (the third mode) are employed for classification. We apply the k-nearest neighbor (k-NN, $k = 1, 3, 5$) classifiers to identify the testing samples in the projected trained sample representation. The classification accuracy, which is defined as the portion of correctly identified entries, with respect to different values of s is reported in Table 6. The results in Table 6 show that classification based on our nonnegative low rank approximation is better than those for other methods in our comparison.

5 Conclusion

In this paper, we proposed a new idea for computing a nonnegative low rank tensor approximation. We proposed a method called NLRT, which determines a nonnegative

[11] Data available at https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html.

Table 5 The number of label samples in each class

No.	1	2	3	4	5	6	7	8
Name	Alfalfa	Corn-no till	Corn-min till	Corn	Grass-pasture	Grass-trees	Grass-pasture-mowed	Hay-windrowed
Samples	46	1428	830	237	483	730	28	478
No.	9	10	11	12	13	14	15	16
Name	Oat	Soybean-no till	Soybean-min till	Soybean-clean	Wheat	Woods	Buildings-Grass-Trees-Drives	Stone-Steel-Towers
Samples	20	972	2455	593	205	1265	386	93

Table 6 The accuracy (in terms of percentage) of the classification results on the approximations by different methods. The **best** values are highlighted in bold

s	Classifier	NCPD-MU	HALS	APG	BCD	CDTF	SaCD	NTD-MU	HALS	APG	BCD	NLRT
10	1-NN	69.68	69.71	67.91	66.40	65.56	61.50	65.89	71.12	73.98	73.70	**74.92**
	3-NN	63.79	64.72	61.89	61.52	60.57	58.00	61.65	65.25	69.80	68.02	**70.12**
	5-NN	62.11	62.72	60.46	60.23	59.21	56.58	61.26	63.67	67.53	65.68	**68.38**
20	1-NN	77.04	77.35	75.05	74.78	74.74	67.95	73.14	79.21	81.16	81.51	**82.06**
	3-NN	72.09	72.39	70.59	69.80	69.53	63.76	69.15	75.20	77.45	76.69	**77.47**
	5-NN	69.59	70.10	68.31	68.54	67.60	63.43	67.53	73.16	75.12	74.55	**75.60**
30	1-NN	81.20	81.01	78.82	79.28	78.36	71.36	76.76	83.19	84.24	85.03	**85.71**
	3-NN	76.76	76.84	74.44	74.37	73.95	68.13	72.11	78.68	80.12	80.91	**81.62**
	5-NN	74.06	74.52	72.38	72.21	72.14	66.46	71.18	76.54	78.29	78.74	**79.16**
40	1-NN	84.19	84.32	81.78	82.09	82.01	74.79	79.38	86.36	86.80	87.18	**88.51**
	3-NN	80.17	79.96	77.99	77.99	78.14	71.17	75.49	81.76	83.93	84.11	**84.87**
	5-NN	78.09	78.34	76.14	75.82	75.87	69.84	74.40	79.80	81.73	81.94	**82.98**
50	1-NN	85.73	86.27	83.50	83.89	83.81	77.15	82.14	88.09	88.16	88.81	**90.19**
	3-NN	82.31	82.14	79.91	80.23	80.42	73.95	78.10	83.95	85.98	85.96	**86.52**
	5-NN	80.19	80.60	77.94	78.32	78.12	72.21	76.95	81.92	84.05	84.03	**84.79**

low rank approximation for the given data by making use of low rank matrix manifolds and the nonnegativity property. A convergence analysis is provided. Experiments with synthetic data sets and multidimensional image data sets are conducted to present the performance of the proposed NLRT method. They show that the NLRT method is better than the classical nonnegative tensor factorization methods.

Acknowledgements T.-X. Jiang's research is supported in part by the National Natural Science Foundation of China under Grant 12001446, the Natural Science Foundation of Sichuan, China under Grant 2022NSFSC1798, and the Fundamental Research Funds for the Central Universities under Grants JBK2202049 and JBK2102001. M. K. Ng's research is supported in part by Hong Kong Research Grant Council GRF 12300218, 12300519, 17201020, 17300021, C1013-21GF, C7004-21GF and Joint NSFC-RGC N-HKU76921. G.-J. Song's research is supported in part by the National Natural Science Foundation of China under Grant 12171369 and Key NSF of Shandong Province under Grant ZR2020KA008.

References

1. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
2. Balasubramaniam, T., Nayak, R., Yuen, C.: Efficient nonnegative tensor factorization via saturating coordinate descent. ACM Trans. Knowl. Discov. Data (TKDD) **14**(4), 1–28 (2020)
3. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and the method of alternating projections: applications. Set Value Var. Anal. **21**(3), 475–501 (2013)
4. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted normal cones and the method of alternating projections: theory. Set Value Var. Anal. **21**(3), 431–473 (2013)
5. Cichocki, A., Phan, A.H.: Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **92**(3), 708–721 (2009)
6. Cichocki, A., Zdunek, R., Amari, S.i.: Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In: International conference on independent component analysis and signal separation, pp. 169–176. Springer (2007)
7. Clarke, F., Vinter, R.: Regularity properties of optimal controls. SIAM J. Control Optim. **28**(4), 980–997 (1990)
8. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
9. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. SIAM J. Matrix Anal. Appl. **21**(4), 1324–1342 (2000)
10. Drusvyatskiy, D., Ioffe, A., Lewis, A.: Alternating projections and coupling slope. arXiv preprint arXiv:1401.7569 pp. 1–17 (2014)
11. Golub, G.H., Van Loan, C.F.: Matrix Computations, vol. 3. JHU Press, Baltimore (2012)
12. Kim, Y.D., Choi, S.: Nonnegative Tucker decomposition. In: 2007 IEEE conference on computer vision and pattern recognition, pp. 1–8. IEEE (2007)
13. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
14. Kroonenberg, P.M.: Applied Multiway Data Analysis, vol. 702. Wiley, New York (2008)
15. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence for alternating and averaged nonconvex projections. Found. Comput. Math. **9**(4), 485–513 (2009)
16. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence for alternating and averaged nonconvex projections. Found. Comput. Math. **9**(4), 485–513 (2009)
17. Lewis, A.S., Malick, J.: Alternating projections on manifolds. Math. Oper. Res. **33**(1), 216–234 (2008)
18. 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)
19. Li, X., Ng, M.K., Cong, G., Ye, Y., Wu, Q.: MR-NTD: manifold regularization nonnegative tucker decomposition for tensor data dimension reduction and representation. IEEE Trans. Neural Netw. Learn. Syst. **28**(8), 1787–1800 (2016)

20. Noll, D., Rondepierre, A.: On local convergence of the method of alternating projections. Found. Comput. Math. **16**(2), 425–455 (2016)
21. Pan, J., Ng, M.K., Liu, Y., Zhang, X., Yan, H.: Orthogonal nonnegative tucker decomposition. arXiv preprint arXiv:1912.06836 (2019)
22. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, vol. 317. Springer, Cham (2009)
23. Shin, K., Sael, L., Kang, U.: Fully scalable methods for distributed tensor factorization. IEEE Trans. Knowl. Data Eng. **29**(1), 100–113 (2016)
24. 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)
25. Song, G.J., Ng, M.K.: Nonnegative low rank matrix approximation for nonnegative matrices. Appl. Math. Lett. (2020). https://doi.org/10.1016/j.aml.2020.106300
26. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
28. Welling, M., Weber, M.: Positive tensor factorization. Pattern Recognit. Lett. **22**(12), 1255–1261 (2001)
29. Xu, Y., Hao, R., Yin, W., Su, Z.: Parallel matrix factorization for low-rank tensor completion. Inverse Probl. Imaging **9**(2), 601–624 (2015)
30. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM J. Imaging Sci. **6**(3), 1758–1789 (2013)
31. Zhang, Y., Zhou, G., Zhao, Q., Cichocki, A., Wang, X.: Fast nonnegative tensor factorization based on accelerated proximal gradient and low-rank approximation. Neurocomputing **198**, 148–154 (2016)
32. Zhou, G., Cichocki, A., Xie, S.: Fast nonnegative matrix/tensor factorization based on low-rank approximation. IEEE Trans. Signal Process. **60**(6), 2928–2940 (2012)