ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096076

VOLUME-REGULARIZED NONNEGATIVE TUCKER DECOMPOSITION WITH IDENTIFIABILITY GUARANTEES

Yuchen Sun and Kejun Huang

Department of CISE, University of Florida, Gainesville, FL 32611
(yuchen.sun,kejun.huang)@ufl.edu

ABSTRACT

It is well-known that the Tucker decomposition of a multi-dimensional tensor is not unique, because its factors are subject to rotation ambiguities similar to matrix factorization models. Inspired by the recent success in the identifiability of nonnegative matrix factorization, the goal of this work is to achieve similar results for nonnegative Tucker decomposition (NTD). We propose to add a matrix volume regularization as the identifiability criterion, and show that NTD is indeed identifiable if all of the Tucker factors satisfy the sufficiently scattered condition. We then derive an algorithm to solve the modified formulation of NTD that minimizes the generalized Kullback-Leibler divergence of the approximation plus the proposed matrix volume regularization. Numerical experiments show the effectiveness of the proposed method.

1. INTRODUCTION

Tensors are multi-dimensional extensions of matrices [1, 2]. The Tucker decomposition [3] of a multiway tensor is perhaps the most natural generalization of the celebrated matrix principal component analysis (PCA) due to its close relationship with the higher-order singular value decomposition (HOSVD) [4, 5]. However, it also inherits the biggest shortcomings of PCA, namely the latent factors are not identifiable due to the inherent rotation ambiguity (without additional constraints on the latent factors). For this reason, the Tucker decomposition is most commonly used as a compression technique rather than an unsupervised factor analysis approach or blind source separation method, unlike most other tensor decomposition models such as the canonical polyadic decomposition [6–10].

Inspired by the success of nonnegative matrix factorization (NMF) [11], there have been nonnegative variants of Tucker decomposition as well [12–15]. Although general matrix factorization is not unique, NMF has been observed to be able to (sometimes, not always) correctly identify the latent factors by simply adding nonnegativity constraints. The most general result to date is that NMF is unique when the latent factors satisfy the 'sufficiently scattered' condition [16]. Furthermore, one could enforce the factors to be sufficiently scattered by optimizing a matrix volume criterion [17]. An overview on identifiability and applications of NMF can be found in [18].

In this paper, we make the first ever attempt to extend such identifiability results to nonnegative tensor decomposition (NTD). We will show that NTD is identifiable, up to scaling and permutation ambiguity, if all the factor matrices are sufficiently scattered. Identifiability of NTD is accomplished by optimizing a novel volume criterion imposed on the core tensor. When noise is present, this naturally leads to a volume-regularized NTF model that jointly fits the data and also uniquely identifies the latent factors. Experiments on synthetic and real data validates the effectiveness of our model.

Supported in part by NSF ECCS-2237640 and NIH R01LM014027.

1.1. Tensors and notations

We denote the input N-way tensor, of size $I_1 \times I_2 \times \cdots \times I_N$, as $\mathcal{X}$. In general, we denote tensors by boldface Euler script capital letters, e.g., $\mathcal{X}$ and $\mathcal{Y}$, while matrices and vectors are denoted by boldface italic capital letters (e.g., $\boldsymbol{X}$ and $\boldsymbol{Y}$) and boldface italic lowercase letters (e.g., $\boldsymbol{x}$ and $\boldsymbol{y}$), respectively. The Euclidean norm of a tensor $\mathcal{X}$ is denoted as $\|\mathcal{X}\|$, which is defined as

$$\|\mathcal{X}\| = \sqrt{\sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \mathcal{X}(i_1, ..., i_N)^2}.$$

Unfolding. A tensor can be unfolded, or *matricized*, along any of its mode into a matrix. The tensor unfolding along the nth mode is denoted $\boldsymbol{X}_{(n)} \in \mathbb{R}^{I_n \times \prod_{v \neq n} I_v}$. More simply, the nth mode of $\mathcal{X}$ forms the rows of $\boldsymbol{X}_{(n)}$ and the remaining modes form the columns.

Tensor-matrix product. The n-mode tensor-matrix product multiplies a tensor with a matrix along the nth mode. Suppose $\boldsymbol{B}$ is a $K \times I_n$ matrix, the n-mode tensor-matrix product, denoted as $\mathcal{X} \times_n \boldsymbol{B}$, outputs a tensor of size $I_1 \times \cdots \times I_{n-1} \times K \times I_{n+1} \times \cdots \times I_N$. Elementwise,

$$\left[\mathcal{X} \times_n \boldsymbol{B}\right](i_1, ..., i_{n-1}, k, i_{n+1}, ..., i_N) = \sum_{i_n=1}^{I_N} \boldsymbol{B}(k, i_n) \mathcal{X}(i_1, ..., i_N).$$

Using mode-n unfolding, it can be equivalently written as

$$\left[\mathcal{X} \times_n \boldsymbol{B}\right]_{(n)} = \boldsymbol{B}\boldsymbol{X}_{(n)}.$$

Note that the resulting tensor is in general dense regardless of the sparsity pattern of $\mathcal{X}$.

A common task is to multiply a tensor by a set of matrices. This operation is called the tensor-times-matrix chain (TTMc). When multiplication is performed with all N modes, it is denoted as $\mathcal{X} \times \{\boldsymbol{B}\}$, where $\{\boldsymbol{B}\}$ is the set of N matrices $\boldsymbol{B}^{(1)}, ..., \boldsymbol{B}^{(N)}$. Sometimes the multiplication is performed with all modes *except one*. This is denoted as $\mathcal{X} \times_{-n} \{\boldsymbol{B}\}$, where n is the mode not being multiplied:

$$\mathcal{X} \times_{-n} \{\boldsymbol{B}\} = \mathcal{X} \times_1 \boldsymbol{B}^{(1)} \cdots \times_{n-1} \boldsymbol{B}^{(n-1)} \times_{n+1} \boldsymbol{B}^{(n+1)} \cdots \times_N \boldsymbol{B}^{(N)}.$$

Kronecker product. The Kroneckder product (KP) of $\boldsymbol{A} \in \mathbb{R}^{\ell \times m}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times q}$, denoted as $\boldsymbol{A} \otimes \boldsymbol{B}$, is an $\ell p \times mq$ matrix defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}(1,1)\boldsymbol{B} & \cdots & \boldsymbol{A}(1,m)\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}(\ell,1)\boldsymbol{B} & \cdots & \boldsymbol{A}(\ell,m)\boldsymbol{B} \end{bmatrix}.$$

Table 1: List of notations

Symbol	Definition
N	number of modes
$\mathcal{X}$	N-way data tensor of size $I_1 \times I_2 \times \cdots \times I_N$
$\mathcal{X}(i_1, ..., i_N)$	$(i_1, ..., i_N)$-th entry of $\mathcal{X}$
$\boldsymbol{X}_{(n)}$	mode-n matrix unfolding of $\mathcal{X}$
I_n	dimension of the nth mode of $\mathcal{X}$
K_n	multilinear rank of the nth mode
$\mathcal{G}$	core tensor of the Tucker model $\in \mathbb{R}^{K_1 \times ... \times K_N}$
$\boldsymbol{U}^{(n)}$	mode-n factor of the Tucker model $\in \mathbb{R}^{I_n \times K_n}$
$\{\boldsymbol{U}\}$	set of all factors $\{\boldsymbol{U}^{(1)}, ..., \boldsymbol{U}^{(N)}\}$
$\times_n$	n-mode tensor-matrix product
$\times_{-n}$	chain of mode products except the nth one

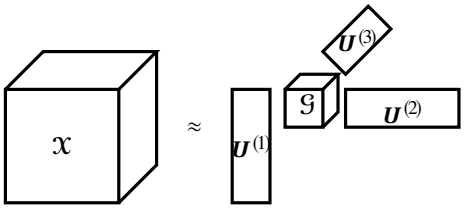


Fig. 1: Tucker decomposition of a 3-way tensor

Mathematically, the n-mode TTMc can be equivalently written as the product of mode-n unfolding times a chain of Kronecker products:

$$\begin{aligned}&\left[\mathcal{X} \times_{-n} \{\boldsymbol{B}\}\right]_{(n)} \\ &= \boldsymbol{X}_{(n)} \left(\boldsymbol{B}^{(1)} \otimes \cdots \otimes \boldsymbol{B}^{(n-1)} \otimes \boldsymbol{B}^{(n+1)} \otimes \cdots \otimes \boldsymbol{B}^{(N)}\right)^{\top}. \end{aligned} \quad (1)$$

More notations are shown in Table 1.

1.2. Nonnegative Tucker decomposition (NTD)

The goal of Tucker decomposition is to approximate a data tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with the product of a core tensor $\mathcal{G} \in \mathbb{R}^{K_1 \times \cdots \times K_N}$ and a set of N factor matrices $\boldsymbol{U}^{(n)} \in \mathbb{R}^{I_n \times K_n}, n = 1, ..., N$, i.e., $\mathcal{X} \approx \mathcal{G} \times \{\boldsymbol{U}\}$. An illustration of Tucker decomposition for 3-way tensors is shown in Figure 1. To find the Tucker decomposition of a given tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ with a target reduced dimension $K_1 \times \cdots \times K_N$, one formulates the following problem:

$$\underset{\substack{\mathcal{G} \in \mathbb{R}^{K_1 \times \cdots \times K_N} \\ \{\boldsymbol{U}^{(n)} \in \mathbb{R}^{I_n \times K_n}\}_{n=1}^{N}}}{\text{minimize}} \quad \|\mathcal{X} - \mathcal{G} \times \{\boldsymbol{U}\}\|^2. \quad (2)$$

Similar to matrix factorization models, the Tucker decomposition suffers from rotation ambiguities: if each factor matrix $\boldsymbol{U}^{(n)}$ is multiplied by a nonsingular matrix $\boldsymbol{A}^{(n)}$ from the left $\boldsymbol{A}^{(n)}\boldsymbol{U}^{(n)}$, the oblique rotation can be 'absorbed' into the core tensor $\mathcal{G}$ as $\mathcal{G} \times \{\boldsymbol{A}^{-1}\}$, which will not affect the overall product

$$\mathcal{G} \times \{\boldsymbol{U}\} = \left(\mathcal{G} \times \{\boldsymbol{A}^{-1}\}\right) \times \{\boldsymbol{U}\boldsymbol{A}\}. \quad (3)$$

For this reason, it is often without loss of generality assumed that the factor matrices all have orthonormal columns. With this constraint, one can eliminate variable $\mathcal{G}$ since it should be equal to $\mathcal{G} = \mathcal{X} \times \{\boldsymbol{U}^{\top}\}$, and equivalently maximize $\|\mathcal{X} \times \{\boldsymbol{U}^{\top}\}\|^2$. A well-known algorithmic framework to approximately optimize it is the higher-order orthogonal iteration (HOOI) [5], which cyclically updates the factors as the K_n leading left singular vectors of $\boldsymbol{Y}_{(n)}$, obtained by taking the n-mode unfolding of the tensor $\mathcal{Y} \triangleq \mathcal{X} \times_{-n} \{\boldsymbol{U}^{\top}\}$. More recently, a novel higher-order QR iteration (HOQRI) was proposed to update the factors as an orthonormal basis of the columns of $\boldsymbol{Y}_{(n)}\boldsymbol{G}_{(n)}^{\top}$, where $\boldsymbol{G}_{(n)}$ is the mode-n unfolding of the core tensor $\mathcal{G}$; the orthonormal basis is usually obtained from the QR factorization [19]. Compared to HOOI, HOQRI avoids the intermediate memory explosion when dealing with large and sparse data tensors (by defining a special kernel to directly calculate $\boldsymbol{Y}_{(n)}\boldsymbol{G}_{(n)}^{\top}$), and is the first Tucker algorithm that is shown to converge to a stationary point.

Nonnegative variants of Tucker decomposition have been proposed in recent years [12] by constraining the variables in (2) to be element-wise nonnegative. However, most of them focus on algorithm designs and not model correctness of why it is beneficial to impose the latent constraints [13, 14]; this question was briefly discussed in [13] and the conclusion was that the latent factors can be uniquely recovered, up to scaling and permutation ambiguity, if they satisfy the separability assumption [20], which is not very realistic in practice. In this paper, we will present a new identifiability result based on the much more practical sufficiently scattered condition [16, 21], and also propose a new algorithm based on Frank-Wolfe.

2. VOLUME REGULARIZED NTD

In this section, we introduce a novel volume criterion into the nonnegative Tucker decomposition, and show that it is able to guarantee unique recovery of the ground-truth latent factors if they satisfy the sufficiently scattered condition, up to scaling and permutation ambiguity. Then we introduce a Frank-Wolfe algorithm based on the formulation of fitting an NTD model with the proposed volume criterion as a regularization.

2.1. Identifiability in the noiseless case

We start by assuming the data tensor $\mathcal{X}$ is generated exactly, without noise, from the Tucker model $\mathcal{G} \times \{\boldsymbol{U}\}$ with nonnegative factors $\boldsymbol{U}^{(n)} \geq 0$ for $n = 1, \ldots, N$. Like all latent variable models, there exist inherent (and inconsequential) scaling and permutation ambiguity regarding the identifiability of the latent factors. Therefore, we define the identifiability of the Tucker factors as follows:

Definition 1 (Identifiability)**.** Consider a data tensor generated from the Tucker model $\mathcal{X} = \mathcal{G}_{\natural} \times \{\boldsymbol{U}_{\natural}\}$, where $\boldsymbol{U}_{\natural}^{(n)} \geq 0, n = 1, \ldots, N$ are the ground-truth factors. Let $\mathcal{G}_{\star}$ and $\{\boldsymbol{U}_{\star}\}$ be optimal for an identification criterion q

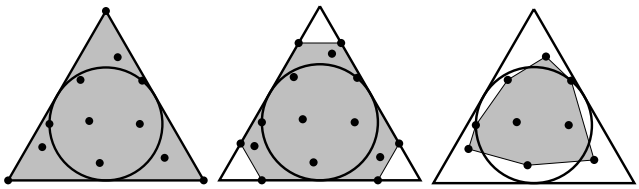
$$(\mathcal{G}_{\star}, \{\boldsymbol{U}_{\star}\}) = \underset{\mathcal{X} = \mathcal{G} \times \{\boldsymbol{U}\}}{\arg\min} \; q(\mathcal{G}, \{\boldsymbol{U}\}).$$

If $\mathcal{G}_{\natural}$ and/or $\{\boldsymbol{U}_{\natural}\}$ satisfy some condition such that, for any $(\mathcal{G}_{\star}, \{\boldsymbol{U}_{\star}\})$, there exist permutation matrices $\boldsymbol{\Pi}^{(1)}, \ldots, \boldsymbol{\Pi}^{(N)}$ and diagonal matrices $\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(N)}$ such that

$$\boldsymbol{U}_{\natural}^{(n)} = \boldsymbol{U}_{\star}^{(n)}\boldsymbol{D}^{(n)}\boldsymbol{\Pi}^{(n)}, \quad n = 1, \ldots, N, \text{ and } \mathcal{G}_{\natural} = \mathcal{G}_{\star} \times \{\boldsymbol{\Pi}^{\top}\boldsymbol{D}^{-1}\},$$

then we say that the NTD model is identifiable under that condition.

Due to the scaling ambiguity and the fact that factor matrices are element-wise nonnegative, it is without loss of generality to assume that each column sums to one, i.e., $\boldsymbol{I}^{\top}\boldsymbol{U}^{(n)} = \boldsymbol{I}^{\top}$, for $n = 1, \ldots, N$. Obviously, this is far from enough to guarantee uniqueness of NTD. Inspired by the recent success of identifiability-guaranteed NMF with



(a) Separable (b) Sufficiently scattered (c) Not identifiable

Fig. 2: A geometric illustration of the sufficiently scattered condition (middle), a special case that is separable (left), and a case that is not identifiable (right). The triangle denotes the nonnegative orthant, the circle denotes the hyperbolic cone C defined in Assumption 1, solid dots represent rows of $\boldsymbol{H}$, and the shaded regions represent cone($\boldsymbol{H}$).

a volume regularization [17, 21], we propose to seek for, among all admissible NTDs, the one that maximizes the volume of each Tucker factor, leading to the following identifiability criterion

$$\begin{aligned} \underset{\substack{\mathcal{G}\in\mathbb{R}^{K_1\times\cdots\times K_N}\\ \{\boldsymbol{U}^{(n)}\in\mathbb{R}^{I_n\times K_n}\}_{n=1}^N}}{\text{maximize}} \quad & \sum_{n=1}^{N} \log\det(\boldsymbol{U}^{(n)\top}\boldsymbol{U}^{(n)}) \\ \text{subject to} \quad & \boldsymbol{U}^{(n)} \geq 0,\ \boldsymbol{I}^\top\boldsymbol{U}^{(n)} = \boldsymbol{I}^\top,\ n = 1,\ldots,N, \\ & \mathcal{X} = \mathcal{G}\times\{\boldsymbol{U}\}. \end{aligned} \tag{4}$$

The determinant of the Gram matrix of a general rectangular matrix is called the volume of a matrix [22]; in this case this is the identification criterion q mentioned in Definition 1. One may notice that, as a new formulation for NTD, (4) does not even include a nonnegativity constrain on the core tensor $\mathcal{G}$. As we will show soon, after removing the nonnegativity constraint, the matrix volume criterion is enough to guarantee identifiability, which makes the decomposition more general by allowing the core tensor to include negative values; if it turns out the core tensor is indeed element-wise nonnegative, identifiability guarantees that it would be exactly recovered (up to permutation and scaling along each mode) even without enforcing the nonnegativity constraint on the core tensor.

The condition that guarantees identifiability of NTD is the sufficiently scattered condition that first appeared in [16] and was further studied in [17, 21] and many others:

Assumption 1 (Sufficiently scattered)**.** Let C denote the hyperbolic cone $\{\boldsymbol{x} \in \mathbb{R}^K \mid \sqrt{K-1}\|\boldsymbol{x}\| \leq \boldsymbol{1}^\top\boldsymbol{x}\}$ and cone($\boldsymbol{H}$) denote the conic hull of the rows of $\boldsymbol{H}$: $\{\boldsymbol{H}^\top\boldsymbol{\theta} \mid \boldsymbol{\theta} \geq 0\}$. A nonnegative matrix $\boldsymbol{H}$ is sufficiently scattered if:

1. $C \subseteq$ cone($\boldsymbol{H}$);
2. $\partial C \cap \partial\,$cone$(\boldsymbol{H}) = \{\alpha(\boldsymbol{1} - \boldsymbol{e}_k) \mid \alpha \geq 0, k = 1,\ldots,K\}$, where ∂ denotes the boundary of the set.

A geometric illustration of a matrix that satisfies the sufficiently scattered condition is shown in Figure 2b, where rows of the matrix are depicted as dots. As we can see, C is a subset of the nonnegative orthant $\mathbb{R}^K_+$, but touches the boundary of $\mathbb{R}^K_+$ at lines $\alpha(\boldsymbol{1} - \boldsymbol{e}_k)$, $k = 1, ..., K$. If a matrix $\boldsymbol{H}$ is sufficiently scattered, cone($\boldsymbol{H}$) contains C as a subset and, as a second requirement, C touches the boundary of cone($\boldsymbol{H}$) only at those points too.

One can also see from Figure 2a that the separability assumption, considered in [20] and in the context of NTD [13], is a very special case of sufficiently scattered. It requires that all the coordinate vectors be included in rows of $\boldsymbol{H}_{\natural}$, which makes cone($\boldsymbol{H}_{\natural}$) $= \mathbb{R}^K_+$, while the sufficiently scattered condition is allowed to grossly violate separability. In fact, it has been empirically observed that a nonnegative sparse matrix satisfies the sufficiently scattered condition with very high probability [18].

Our main result on the identifiability of NTD is presented as follows:

Theorem 1. *Assume that $\mathcal{X} = \mathcal{G}_{\natural} \times \{\boldsymbol{U}_{\natural}\}$, where all the ground-truth nonnegative Tucker factors $\boldsymbol{U}^{(n)}_{\natural}$ are sufficiently scattered (Assumption 1). Let $(\mathcal{G}_\star, \{\boldsymbol{U}_\star\})$ be an optimal solution of* (4)*, then there exist permutation matrices $\boldsymbol{\Pi}^{(1)}, \ldots, \boldsymbol{\Pi}^{(N)}$ and diagonal matrices $\boldsymbol{D}^{(1)}, \ldots, \boldsymbol{D}^{(N)}$ such that*

$$\boldsymbol{U}^{(n)}_{\natural} = \boldsymbol{U}^{(n)}_\star \boldsymbol{D}^{(n)}\boldsymbol{\Pi}^{(n)}, \quad n = 1,\ldots,N, \ \text{and}\ \ \mathcal{G}_{\natural} = \mathcal{G}_\star \times \{\boldsymbol{\Pi}^\top\boldsymbol{D}^{-1}\}.$$

In other words, NTD is identifiable (Definition 1) if all the Tucker factors are sufficiently scattered.

Due to space limitation, the proof is relegated to the journal version.

2.2. Algorithm

In practice, the data tensor most likely does not admit an exact NTD $\mathcal{X} = \mathcal{G} \times \{\boldsymbol{U}\}$. Therefore, when designing an algorithm for identifiability guaranteed NTD, one has to balance the identification criterion, the volumes of the Tucker factors in this case, and data fidelity. We propose to formulate the problem as

$$\begin{aligned} \underset{\substack{\mathcal{G}\in\mathbb{R}^{K_1\times\cdots\times K_N}\\ \{\boldsymbol{U}^{(n)}\in\mathbb{R}^{I_n\times K_n}\}_{n=1}^N}}{\text{minimize}} \quad & D(\mathcal{X}\|\mathcal{G}\times\{\boldsymbol{U}\}) - \lambda\sum_{n=1}^{N}\log\det(\boldsymbol{U}^{(n)\top}\boldsymbol{U}^{(n)}) \\ \text{subject to} \quad & \mathcal{G} \geq 0,\ \boldsymbol{U}^{(n)} \geq 0,\ \boldsymbol{I}^\top\boldsymbol{U}^{(n)} = \boldsymbol{I}^\top,\ n = 1,\ldots,N, \end{aligned} \tag{5}$$

where λ is the regularization parameter that controls the balance between data fidelity and the identification criterion, and $D(\cdot\|\cdot)$ is the generalized Kullback-Leibler (GKL) divergence defined as

$$D(\mathcal{X}\|\mathcal{G}\times\{\boldsymbol{U}\}) = \sum_{i_1,\ldots,i_N}\left(\mathcal{X}(i_1,\ldots,i_N)\log\frac{\mathcal{X}(i_1,\ldots,i_N)}{\mathcal{G}\times\{\boldsymbol{u}^{(1)}_{i_1},\ldots,\boldsymbol{u}^{(N)}_{i_N}\}} - \mathcal{X}(i_1,\ldots,i_N) + \mathcal{G}\times\{\boldsymbol{u}^{(1)}_{i_1},\ldots,\boldsymbol{u}^{(N)}_{i_N}\}\right).$$

Ignoring terms that do not depend on the variables, and using the fact that columns of $\boldsymbol{U}^{(n)}$ all sum to one, the GKL divergence is equivalent to (up to a constant difference)

$$\sum\mathcal{G} - \sum\left(\mathcal{X} * \log(\mathcal{G}\times\{\boldsymbol{U}\})\right), \tag{6}$$

where we overload the notation $\sum$ to denote summation over all elements of the tensor, $*$ denote element-wise multiplication, and the log of a tensor is also taken element-wise.

Since Problem 5 is non-convex, we propose to approximately solve it using successive convex approximation (SCA) [23]. At iteration t when the updates are $\mathcal{G}_t$ and $\{\boldsymbol{U}_t\}$, we define

$$\begin{aligned} &\Pi_t(i_1,\ldots,i_N,k_1,\ldots,k_N) \\ &= \frac{\mathcal{G}_t(k_1,\ldots,k_N)\boldsymbol{U}^{(1)}_t(i_1,k_1)\cdots\boldsymbol{U}^{(N)}_t(i_N,k_N)}{\sum_{k_1,\ldots,k_N}\mathcal{G}_t(k_1,\ldots,k_N)\boldsymbol{U}^{(1)}_t(i_1,k_1)\cdots\boldsymbol{U}^{(N)}_t(i_N,k_N)}. \end{aligned}$$

Obviously $\sum_{k_1,\ldots,k_N} \Pi_t(i_1,\ldots,i_N,k_1,\ldots,k_N) = 1$ and $\Pi_t(i_1,\ldots,i_N,k_1,\ldots,k_N) \geq 0$, which defines a probability mass function for each $(i_1,\ldots,i_N)$. Using Jensen's inequality, we have that

$$
\begin{aligned}
&-\mathcal{X}(i_1,\ldots,i_N)\log\sum_{k_1,\ldots,k_N}\mathcal{G}(k_1,\ldots,k_N)\boldsymbol{U}^{(1)}(i_1,k_1)\cdots\boldsymbol{U}^{(N)}(i_N,k_N)\\
&\leq -\sum_{k_1,\ldots,k_N}\mathcal{X}(i_1,\ldots,i_N)\Pi_t(i_1,\ldots,i_N,k_1,\ldots,k_N)\times\\
&\qquad\Big(\log\mathcal{G}(k_1,\ldots,k_N)+\log\boldsymbol{U}^{(1)}(i_1,k_1)+\cdots+\log\boldsymbol{U}^{(N)}(i_N,k_N)\\
&\qquad\qquad-\log\Pi_t(i_1,\ldots,i_N,k_1,\ldots,k_N)\Big),
\end{aligned}
$$

which defines a convex and locally tight upperbound for the first term in the loss function of (5). Regarding the second term, we propose to simply take the linear approximation

$$
\log\det(\boldsymbol{U}^{(n)\mathsf{T}}\boldsymbol{U}^{(n)}) \approx \log\det(\boldsymbol{U}_t^{(n)\mathsf{T}}\boldsymbol{U}_t^{(n)}) + 2\,\mathrm{Tr}\,\boldsymbol{U}_t^{(n)\dagger}\left(\boldsymbol{U}^{(n)}-\boldsymbol{U}_t^{(n)}\right),
$$

where $2(\boldsymbol{U}_t^{(n)\dagger})^\mathsf{T}$ is the gradient of $\log\det(\boldsymbol{U}_t^{(n)\mathsf{T}}\boldsymbol{U}_t^{(n)})$.

Now that we have derived a convex approximation to the objective of (5), which is separable down to each scalar variable, we can obtain the SCA updates without much difficulty. Due to space limitations, we skip some of the tedious steps and directly present the SCA algorithm as in (1). We would like to make two comments: 1) the operation performed in line 5 is mathematically represented as matrix multiplication of the n-mode matricization of $\widetilde{\mathcal{X}}\times_{-n}\{\boldsymbol{U}^\mathsf{T}\}$ and the transpose of that of $\mathcal{G}$; if the data tensor is large and sparse, this operation can be done efficiently via the TTMcTC (stands for tensor times matrix chain times core) kernel without instantiating the large and dense intermediate tensors [19]; and 2) the scalar α in line 13 corresponds to the Lagrange multiplier of the constraint $\boldsymbol{I}^\mathsf{T}\boldsymbol{u}=1$; even though it is the solution of a nonlinear equation that cannot be solved analytically, it can be efficiently computed via bi-section.

Algorithm 1 Proposed algorithm: Solving (5) with SCA

1: initialize $\mathcal{G}$ and $\{\boldsymbol{U}\}$
2: **repeat**
3: $\quad\widetilde{\mathcal{X}} = \mathcal{X}/(\mathcal{G}\times\{\boldsymbol{U}\})$ ▷ element-wise division
4: $\quad$**for** $n = 1,\ldots,N$ **do**
5: $\quad\quad\widetilde{\boldsymbol{U}}^{(n)} \leftarrow [\widetilde{\mathcal{X}}\times_{-n}\{\boldsymbol{U}^\mathsf{T}\}]_{(n)}\boldsymbol{G}_{(n)}^\mathsf{T}$
6: $\quad$**end for**
7: $\quad\mathcal{G}\leftarrow\mathcal{G}*(\widetilde{\mathcal{X}}\times\{\boldsymbol{U}^\mathsf{T}\})$ ▷ element-wise multiplication
8: $\quad$**for** $n = 1,\ldots,N$ **do**
9: $\quad\quad\boldsymbol{V} = 2(\boldsymbol{U}^{(n)\dagger})^\mathsf{T}$
10: $\quad\quad$**for** $k_n = 1,\ldots,K_N$ **do**
11: $\quad\quad\quad$denote $\boldsymbol{v}$ as the kth column of $\boldsymbol{V}$
12: $\quad\quad\quad$denote $\tilde{\boldsymbol{u}}$ as the kth column of $\widetilde{\boldsymbol{U}}^{(n)}$
13: $\quad\quad\quad$find scalar α such that $\boldsymbol{u} = \tilde{\boldsymbol{u}}/(-\lambda\boldsymbol{v}+\alpha) > 0$ and $\boldsymbol{I}^\mathsf{T}\boldsymbol{u}=1$
14: $\quad\quad\quad$update the kth column of $\boldsymbol{U}^{(n)}$ as $\boldsymbol{u}$
15: $\quad\quad$**end for**
16: $\quad$**end for**
17: **until** convergence

Regarding initialization, we propose to start by applying any algorithm for Tucker decomposition with orthonormal constraints, such as HOOI [5] or HOQRI [19], then apply the algorithm in [21] on each factor to obtain an initialization of $\boldsymbol{U}^{(n)}$; the oblique rotations are then absorbed into the core tensor, followed by setting all negative values as zeros as initialization of $\mathcal{G}$.

3. NUMERICAL VALIDATION

We conclude the paper by providing some numerical validation to the proposed theoretical analysis. We focus on 3-way tensors of dimension $I_1 = I_2 = I_3 = 100$ and multilinear ranks $K_1 = K_2 = K_3 = 10$. Since the focus of this paper is identifiability, we will synthetically generate the ground-truth Tucker factors $\boldsymbol{U}_\natural^{(1)}, \boldsymbol{U}_\natural^{(n)}, \boldsymbol{U}_\natural^{(3)}$ and the core tensor $\mathcal{G}_\natural$, multiply them to get the data tensor $\mathcal{X} = \mathcal{G}\times_1\boldsymbol{U}_\natural^{(1)}\times_2\boldsymbol{U}_\natural^{(n)}\times_3\boldsymbol{U}_\natural^{(3)}$, possibly contaminated with some noise. All the positive elements in the ground-truth factors are generated from independent exponential distributions. A portion of randomly selected elements in the ground-truth factors are set to zeros, since it has been observed that a sparse latent factor satisfies the sufficiently scattered condition with very high probability [18]. To resolve the scaling ambiguity, all columns of $\boldsymbol{U}_\natural^{(n)}$ are rescaled to sum to one, leaving only permutation ambiguity to be resolved in the end.

In our first numerical experiment, we vary the level of sparsity of the latent factors and check how it affects identifiability. It has been shown in [16] that if a $I_n\times K_n$ matrix is sufficiently scattered, then each columns of it contains at least K_n-1 zeros. This gives a rule-of-thumb of how sparse the latent factors should be in order to guarantee identifiability. Since we fix $I_n = 100$ and $K_n = 10$, we could expect the model to be identifiable when the density, meaning the percentage of elements being nonzero, is lower than 90%. We vary the latent density from 50% to 95%, and check the probability of exact recovery. In each case, we generate 100 random instances of the ground-truth factors and the core tensor, multiply them to get the data tensor, and apply the initialization strategy of Algorithm 1. After resolving the permutation matrix via the Hungarian algorithm, we declare success if the estimation errors of all of the latent factors are less than 10^{-5}. As we can see, the probability of success remains close to 1 even when the latent density is at the marginal 90%, but quickly goes to zero once it becomes higher.

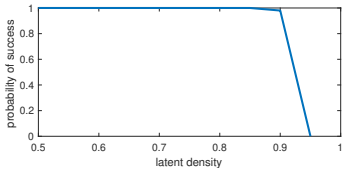


Fig. 3: Probability of exact recovery of the latent factors as we vary the density of the latent factors.

Finally, we demonstrate the convergence behavior of the proposed Algorithm 1. In this case the data tensor $\mathcal{X}$ is no longer noiseless. Since Algorithm 1 tries to solve Problem (5) with the generalized KL divergence, it makes sense to generate the elements of $\mathcal{X}$ from independent Poisson distributions parameterized by the corresponding values in the Tucker product of the ground-truth factors. As we can see in Fig. 4, the algorithm does monotonically decrease the loss value. Due to the Poisson noise, the loss is not close to zero. However, as we will elaborate in the journal paper, the introduced volume-regularization still helps reduce the estimation errors of the latent factors.

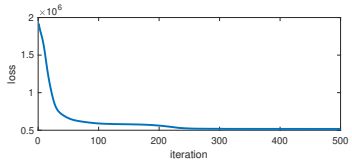


Fig. 4: An instance of the convergence of Algorithm 1.

4. REFERENCES

[1] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[2] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[3] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[4] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[5] ——, "On the best rank-1 and rank-(R_1,R_2,...,R_N) approximation of higher-order tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[6] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927.

[7] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multimodal factor analysis," *UCLA Working Papers Phonetics*, vol. 16, pp. 1–84, 1970.

[8] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, vol. 18, no. 2, pp. 95–138, 1977.

[9] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 14, no. 3, pp. 229–239, 2000.

[10] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE transactions on Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[12] Y.-D. Kim and S. Choi, "Nonnegative Tucker decomposition," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[13] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie, "Efficient nonnegative Tucker decompositions: Algorithms and uniqueness," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4990–5003, 2015.

[14] Y. Xu, "Alternating proximal gradient method for sparse nonnegative Tucker decomposition," *Mathematical Programming Computation*, vol. 7, no. 1, pp. 39–70, 2015.

[15] J. E. Cohen, P. Comon, and N. Gillis, "Some theory on nonnegative Tucker decomposition," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 152–161.

[16] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2013.

[17] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 328–332, 2018.

[18] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 36, no. 2, pp. 59–80, 2019.

[19] Y. Sun and K. Huang, "HOQRI: Higher-Order QR Iteration for Scalable Tucker Decomposition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3648–3652.

[20] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" *Advances in neural information processing systems*, vol. 16, 2003.

[21] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[22] A. Ben-Israel, "A volume associated with $m \times n$ matrices," *Linear Algebra and Its Applications*, vol. 167, pp. 87–111, 1992.

[23] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.