Tensor Principal Component Analysis in High Dimensional CP Models

Yuefeng Han and Cun-Hui Zhang

***Abstract*— The CP decomposition for high dimensional non-orthogonal spiked tensors is an important problem with broad applications across many disciplines. However, previous works with theoretical guarantee typically assume restrictive incoherence conditions on the basis vectors for the CP components. In this paper, we propose new computationally efficient composite PCA and concurrent orthogonalization algorithms for tensor CP decomposition with theoretical guarantees under mild incoherence conditions. The composite PCA applies the principal component or singular value decompositions twice, first to a matrix unfolding of the tensor data to obtain singular vectors and then to the matrix folding of the singular vectors obtained in the first step. It can be used as an initialization for any iterative optimization schemes for the tensor CP decomposition. The concurrent orthogonalization algorithm iteratively estimates the basis vector in each mode of the tensor by simultaneously applying projections to the orthogonal complements of the spaces generated by other CP components in other modes. It is designed to improve the alternating least squares estimator and other forms of the high order orthogonal iteration for tensors with low or moderately high CP ranks, and it is guaranteed to have second or higher order convergence when the error of any given initial estimator is bounded by a small constant. Our theoretical investigation provides estimation accuracy and convergence rates for the two proposed algorithms. Both proposed algorithms are applicable to deterministic tensor, its noisy version, and the order-$2K$ covariance tensor of order-K tensor data in a factor model with uncorrelated factors. Simulation experiments demonstrate significant practical superiority of our approach over existing methods.**

***Index Terms*— Tensor principal component analysis, PCA, CP decomposition, spiked covariance, dimension reduction, unfolding, orthogonal projection.**

I. Introduction

MOTIVATED by modern scientific research, analysis of tensors, or high-order arrays, has emerged as one of the most important and active areas in machine learning, electrical engineering, and statistics. Tensors arise in numerous applications involving genomics [1], [2], multi-relational learning [3], neuroimaging analysis [4], [5], recommender systems [6], computer vision [7], longitudinal data analysis [8], economic indicators [9], [10], finance data [11] and more. In addition, tensor based methods have been applied to many statistics and machine learning problems where the observations are not necessarily tensors, such as community detection [12], topic and latent variable models [13], graphical models [14], and high-order interaction pursuit [15]. In many of these settings, the tensor of interest is high-dimensional, e.g. the ambient dimension is substantially larger than the sample size in the factor model in (1) below. However, in practice, the tensor parameter often has intrinsic dimension-reduced structure, such as low-rankness and sparsity [16], [17], which motivates research in tensor estimation and in the recovery of the underlying structure.

Low rank tensor decomposition is one of the most important tools for recovering and estimating the intrinsic tensor structure based on noisy data. It plays a similar role to matrix singular value decomposition (SVD) and eigendecomposition which are of fundamental importance throughout a wide range of fields including computer science, applied mathematics, machine leaning, statistics, signal processing, etc. Despite the well-established theory for low-rank decomposition of matrices, tensors present unique challenges. There are several notions of low-rankness in tensors, including the most popular CANDECOMP/PARAFAC (CP) low-rankness and multilinear/Tucker low-rankness. While CP models are more parsimonious and easier to interpret in many applications, compared with Tucker models, the computation of the best low-rank CP approximation of a given tensor is NP hard in general [16], [18], [19].

In this paper, we develop a new framework of tensor principal component analysis (tensor PCA) applicable to deterministic tensors, their noisy version, and factor models with uncorrelated factors. To be specific, let us first consider the factor model. Suppose we have i.i.d. matrix or tensor valued observations (such as 2-D or 3-D images) $\mathcal{X}_i$, $1 \leqslant i \leqslant n$, of the following form

$$\mathcal{X}_i = \textstyle\sum_{j=1}^r w_j f_{ij} a_{j1} \otimes a_{j2} \otimes \cdots \otimes a_{jK} + \mathcal{E}_i, \qquad (1)$$

where $\otimes$ denotes tensor product, f_{ij} are i.i.d $N(0,1)$, $w_j > 0$ represent certain weights, $a_{jk} \in \mathbb{R}^{d_k}$ are basis vectors with $\|a_{jk}\|_2 = 1$ for all $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant K$, $\mathcal{E}_i$ are i.i.d. noise tensors each with i.i.d $N(0,\sigma^2)$ entries. Tensor factor models like (1), where the data is written as the sum of a

Manuscript received 12 November 2021; revised 5 May 2022; accepted 7 July 2022. Date of publication 5 September 2022; date of current version 20 January 2023. The work of Yuefeng Han was supported in part by the NSF under Grant IIS-1741390. The work of Cun-Hui Zhang was supported in part by the NSF under Grant IIS-1741390, Grant CCF-1934924, Grant DMS-2052949, and Grant DMS-2210850. An earlier version of this paper was presented in part at the 2022 IMS Annual Meeting and in part at the 2022 NASDS Conference. *(Corresponding author: Cun-Hui Zhang.)*

Yuefeng Han is with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: yuefeng.han@nd.edu).

Cun-Hui Zhang is with the Department of Statistics, Rutgers University, New Brunswick, NJ 08854 USA (e-mail: czhang@stat.rutgers.edu).

Communicated by V. Y. F. Tan, Associate Editor for Machine Learning and Statistics.

Digital Object Identifier 10.1109/TIT.2022.3203972

low-rank factor and noise, have been studied extensively in the literature. While both the Tucker and CP decompositions can be used to model such data, the Tucker model has been the focus of the literature in the consistent estimation of tensor structure in the presence noise, largely due to the direct expression of the Tucker decomposition with matrix SVD. However, in many applications, CP decomposition is a more attractive modeling option. To estimate the structural parameters in model (1), we construct the covariance tensor of the data $\mathcal{X}_i$, $T = n^{-1}\sum_{i=1}^{n}\mathcal{X}_i \otimes \mathcal{X}_i$, which can be written as

$$T = \sum_{j=1}^{r}\lambda_j \otimes_{k=1}^{2K} a_{jk} + \Psi, \tag{2}$$

where $\lambda_j = w_j^2$, $a_{j,K+k} = a_{jk}$ for $1 \leqslant j \leqslant K$, and Ψ is a noise tensor. We treat (2) as a general CP model in which Ψ is allowed to have an identity component $\mathbb{E}[\Psi] \propto \mathrm{Id}$, e.g. $\mathbb{E}[\Psi] = \sigma^2\mathrm{Id}$ under (1), as the estimated basis vectors do not depend on the identity component in our approach. Here Id is the identity tensor given by $\max_{[K]}(\mathrm{Id}) = I_d$. Our main goal is to estimate the basis vectors a_{jk}, which can be also called loading vectors, from the noisy tensor T. We call (2) spiked covariance tensor model as it is analogous to the so called "spiked covariance model" in the study of matrix PCA in high dimensions [20]. Here, a_{jk}, $1 \leqslant j \leqslant r$, are not necessarily orthogonal to each other in each mode k.

When $K = 2$, our model (1) is closely related to $(2D)^2$-PCA in the community of image signal processing, which has been extensively studied [21], [22], [23], [24], [25], [26], [27], [28]. This literature has been mainly focused on the algorithmic properties. However, statistical guarantees such as consistency of estimators and risk analysis, in high demand in many applications, are much less understood in the CP model. Among notable exceptions is [29]. In Tucker factor models, statistical analysis has been carried out by [30], [31], [32], [9], [33], and [34] under a very different setting from (1), and by [35], [10], and [36] with matrix and tensor time series.

In view of (2), a natural approach to the estimation of a_{jk} is minimizing the empirical loss,

$$\underset{\|a_{jk}\|_2=1\forall j,k}{\arg\min}\ \min_{\lambda_j\forall j}\left\|T - \sum_{j=1}^{r}\lambda_j(\otimes_{k=1}^{K}a_{jk})^{\otimes 2}\right\|_{\mathrm{HS}}^2, \tag{3}$$

where $\|\mathcal{A}\|_{\mathrm{HS}}$, defined as $\|\mathrm{vec}(\mathcal{A})\|_2$, is the Hilbert Schmidt norm of tensor $\mathcal{A}$. However, due to the non-convexity of (3), a straightforward implementation of local search algorithms, such as gradient descent and alternating minimization, may get trapped into local optimums and result in sub-optimal statistical performance. Still, if one starts from an initialization not too far from the true basis vectors, local search is likely to perform well.

In addition to the order $2K$ tensor T in (2) with paired CP basis vectors a_{jk}, we study in this paper the following more general low-rank CP model,

$$T = \sum_{j=1}^{r}\lambda_j \otimes_{k=1}^{N} a_{jk} + \Psi, \tag{4}$$

where $\lambda_j > 0$ and Ψ is a noise tensor, including the noiseless version with $\Psi = 0$. While a_{jk} can be all different in (4), $a_{j1} = \cdots = a_{jN}$ for the empirical N-th moment tensor in certain latent-variable models [13].

A. Our Contributions

We propose a new method for the estimation of the basis/loading vectors $a_{jk} \in \mathbb{R}^{d_k}$ in the spiked covariance tensor model (2) or the low-rank CP model (4), both can be viewed as spiked CP models. The new method is composed of two steps: (i) a composite PCA (CPCA) as a warm-start initialization; (ii) an iterative concurrent orthogonalization (ICO) scheme to refine the estimator. The intuition is that the CP components in higher order tensors are closer to orthogonal and tend to have higher order coherence in a multiplicative form, and the proposed method is designed to take advantage of this feature of the CP model to achieve higher statistical and computational efficiency. To the best of our knowledge, this proposal is the first to explicitly aim to benefit from this multiplicative higher order coherence in CP decomposition. Existing initialization procedures require random projections and may need to generate many copies to yield a reasonably good choice, while the CPCA produces definitive initial estimates of the CP basis vectors via tensor unfolding/refolding and spectral decomposition. The ICO scheme aims to achieve higher order of numerical convergence than the alternating least squares and other forms of the high order orthogonal iteration (HOOI) [36], [37], [38], [39] after the warm-start, again by taking benefits of the multiplicative coherence.

The CPCA and ICO algorithms are developed in Section II along with some sharp tensor perturbation bounds to motivate them. These tensor perturbation bounds, new or not readily available and potentially useful elsewhere, heuristically justify our ideas and the individual elements of the proposed algorithms. Statistical guarantees for the CPCA and ICO estimators are provided in Section III. Our perturbation and risk bounds explicitly exhibit the benefits of the multiplicative nature of the coherence of the tensor bases and the rapid growth of such benefits as the order of the tensor increases.

Our analyses of the proposed methods focus on the cases where the tensor dimensions d_k are typically much larger than the CP rank r but r can be also large. Theoretical studies of existing proposals of tensor de-noising in CP models typically imposes very restrictive incoherence conditions on the CP components; For example, the incoherence condition $\vartheta_{\max} = \max_{k,j_1\neq j_2}|a_{j_1k}^\top a_{j_2k}| \lesssim$ ploylog$(d_k)/\sqrt{d_k}$ in [29]. In contrast, we prove that the CPCA yields useful estimates when $r^2\vartheta_{\max}^K$ is small and the ICO provides fast convergence rates when $r^{5/2}\vartheta_{\max}^K$ is small, demonstrating the advantage of our approach in terms of model assumption. Computationally, the errors in the ICO propagate in the quadratic or higher order. Similar to Nesterov's acceleration in gradient descent, the high-order of error propagation guarantees ϵ numerical precision within $\log\log(1/\epsilon)$ iterations. To the best of our knowledge, this is the first provable $\log\log(1/\epsilon)$ iteration guarantee in non-orthogonal CP models. Numerical comparisons with existing methods demonstrate advantages of the proposed approach.

B. Related Work

There is a large literature on tensor decomposition. As it is beyond the scope of this paper to give a comprehensive survey, we only review the most related papers.

The most commonly used algorithm for CP decomposition is alternating least squares [40], which has no general convergence guarantee. Theoretical studies of alternating least squares have focused on the estimation of tensors with orthogonal CP decomposition from randomized initialization. Noticeably, [13] developed a robust tensor power method for orthogonal CP decomposition. [41] proposed a tensor unfolding approach for rank-one tensors and compared it with tensor power iteration with random initialization. [42] improved the initialization procedure for orthogonal CP decomposition by projecting the observed tensor down to a matrix and then applying the matrix power method. [43] developed a two-mode higher-order SVD algorithm for higher order tensors. For the non-orthogonal tensors, one may first convert the tensor into an orthogonal form known as *whitening*, but the procedure is ill-conditioned in high dimensions [44], [45] and computationally expensive [46].

Recently, another line of research has been developed on non-orthogonal tensor CP decomposition, still focused on randomized initialization. [29] studied non-orthogonal CP decomposition and established convergence guarantees for a modification of the alternating least squares. In addition to their incoherence conditions on deterministic CP bases as discussed in the previous subsection, they considered independent random basis vectors uniformly distributed in the unit sphere, essentially imposing a *soft orthogonality* constraint. [47] further extended their work to the case where the CP basis vectors are sparse. In the noiseless case ($\Psi = 0$), [48] introduced orthogonalized alternating least squares algorithm and studied its performance under the soft orthogonality constraint or small $r^2\vartheta_{\max}$. [49] developed a minimum distance algorithm for non-orthogonal CP decomposition which uses random projections to reduce the problem to simultaneous matrix diagonalization, but the applicability of their theoretical results to diverging d_k is unclear. [50] developed an iterative Gauss-Newton algorithm for joint matrix diagonalization. However, [48] claimed that the simultaneous diagonalization algorithm is not computationally efficient.

C. Notation and Tensor Preliminaries

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For a vector with entries π_j or a set of real numbers $\{\pi_j\}$, we denote by $\pi_{j,\pm} = \min_{i\neq j} |\pi_i - \pi_j| \wedge |\pi_j|$ the gap from π_j to $\{0, \pi_i, i \neq j\}$ and set $\pi_{\min} = \min_j |\pi_j|$ and $\pi_{\max} = \max_j |\pi_j|$. For convenience, we call $\lambda_{j,\pm}$ the j-th eigengap in models (2) and (4). For a matrix $B = (b_{ij}) \in \mathbb{R}^{p\times n}$, we denote its singular values by $\sigma_1(B) \geqslant \sigma_2(B) \geqslant \cdots \geqslant \sigma_{\min\{p,n\}}(B) \geqslant 0$, its Frobenius norm by $\|B\|_{\mathrm{F}} = (\sum_{ij} b_{ij}^2)^{1/2} = (\sum_{j=1}^{\min\{p,n\}} \sigma_i^2(B))^{1/2}$, and its spectral norm by $\|B\|_{\mathrm{S}} = \sigma_1(B)$.

For any two vectors u and $\widehat{u}$ of unit length, we measure the distance between the spaces they generate by the absolute sine of the angle $\theta(\widehat{u}, u)$ between the two vectors,

$$\begin{aligned} \big|\sin\theta(\widehat{u}, u)\big| &= \|\widehat{u}\widehat{u}^\top - uu^\top\|_{\mathrm{S}} = (1 - (u^\top\widehat{u})^2)^{1/2} \\ &= \|\widehat{u}\widehat{u}^\top - uu^\top\|_{\mathrm{F}}/\sqrt{2}. \end{aligned} \tag{5}$$

We note that $\min_\pm \|\widehat{u} \pm u\|_2 = (1 - |u^\top\widehat{u}|)^{1/2} = \big|\sin\theta(\widehat{u}, u)\big|/(1 + |u^\top\widehat{u}|)^{1/2}$.

For any two tensors $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}, \mathcal{B} \in \mathbb{R}^{r_1\times r_2\times\cdots\times r_N}$, denote the tensor product $\otimes$ as $\mathcal{A}\otimes\mathcal{B} \in \mathbb{R}^{m_1\times\cdots\times m_K\times r_1\times\cdots\times r_N}$, such that $(\mathcal{A}\otimes\mathcal{B})_{i_1,\ldots,i_K,j_1,\ldots,j_N} = (\mathcal{A})_{i_1,\ldots,i_K}(\mathcal{B})_{j_1,\ldots,j_N}$. For two vectors a and b, $a\otimes b$ is equivalent to the outer product $ab^\top$. Given $\mathcal{A} \in \mathbb{R}^{m_1\times\cdots\times m_K}$ and $m = \prod_{j=1}^K m_j$, let $\mathrm{vec}(\mathcal{A}) \in \mathbb{R}^m$ be vectorization of the matrix/tensor $\mathcal{A}$, $\mathrm{mat}_k(\mathcal{A}) \in \mathbb{R}^{m_k\times(m/m_k)}$ the mode-k matrix unfolding of $\mathcal{A}$, and $\mathrm{mat}_k(\mathrm{vec}(\mathcal{A})) = \mathrm{mat}_k(\mathcal{A})$. For example, for $K = 3$

$$\begin{aligned} (\mathrm{mat}_1(\mathcal{A}))_{i,(j+m_2(k-1))} &= (\mathrm{mat}_2(\mathcal{A}))_{j,(k+m_3(i-1))} \\ &= (\mathrm{mat}_3(\mathcal{A}))_{k,(i+m_1(j-1))} = \mathcal{A}_{ijk}. \end{aligned}$$

Similarly, for nonempty $J \subseteq [K]$, $\mathrm{mat}_J(\mathcal{A})$ is the mode J matrix unfolding which maps $\mathcal{A}$ to $m_J \times m_{-J}$ matrix with $m_J = \prod_{j\in J} m_j$ and $m_{-J} = m/m_J$, e.g. $\mathrm{mat}_{\{1,2\}}(\mathcal{A}) = \mathrm{mat}_3^\top(\mathcal{A})$ for $K = 3$. The mode-k product of $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$ with a matrix $U \in \mathbb{R}^{m_k\times r_k}$ is an order K-tensor of size $m_1 \times \cdots m_{k-1} \times r_k \times m_{k+1} \times m_K$ and will be denoted as $\mathcal{A}\times_k U$, so that

$$(\mathcal{A}\times_k U)_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} = \textstyle\sum_{i_k=1}^{m_k} \mathcal{A}_{i_1,i_2,\ldots,i_K} U_{i_k,j}.$$

The Hilbert Schmidt norm for a tensor $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$ is defined as $\|\mathcal{A}\|_{\mathrm{HS}} = \|\mathrm{vec}(\mathcal{A})\|_2$. An order K tensor $T \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$ is said to have rank one if it can be written as

$$T = w \cdot a_1 \otimes \cdots \otimes a_K,$$

where $w \in \mathbb{R}$ and $a_k \in \mathbb{R}^{m_k}$ are unit vectors for identifiability. A tensor $T \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$ is said to have a CP rank $r \geqslant 1$ if it can be written as a sum of r rank-1 tensors,

$$T = \textstyle\sum_{j=1}^r w_j \cdot a_{i1} \otimes \cdots \otimes a_{iK}.$$

II. Estimation Procedures

A. Spiked Covariance Tensor Model

In this section, we focus on the spiked covariance tensor model (2). We introduce the composite PCA (CPCA) as Algorithm 1, and the iterative concurrent orthogonalization (ICO) as Algorithm 2.

As mentioned in the introduction, our main idea is to take advantage of the multiplicative higher-order coherence of the CP components for faster convergence. We begin with an explicit description of this phenomenon. Let $\Sigma_k = (\sigma_{ij,k})_{r\times r} = A_k^\top A_k$ with the mode-k basis matrix $A_k = (a_{1k}, \ldots, a_{rk}) \in \mathbb{R}^{d_k\times r}$ in (2). As $\sigma_{jj,k} = \|a_{jk}\|_2^2 = 1$, the correlation among columns of A_k can be measured by

$$\begin{aligned} &\vartheta_k = \max\nolimits_{1\leqslant i<j\leqslant r} |\sigma_{ij,k}|, \quad \delta_k = \|\Sigma_k - I_r\|_{\mathrm{S}}, \\ &\eta_{jk} = (\textstyle\sum_{i\in[r]\setminus\{j\}} \sigma_{ij,k}^2)^{1/2}. \end{aligned} \tag{6}$$

However, the CP components are much less correlated. By (2), the matrix unfolding of T,

$$\begin{aligned}\mathrm{mat}_{[K]}(T) &= n^{-1}\textstyle\sum_{i=1}^{n}\mathrm{vec}(\mathcal{X}_i)\mathrm{vec}(\mathcal{X}_i)^\top \\ &= \textstyle\sum_{j=1}^{r}\lambda_j a_j a_j^\top + \mathrm{mat}_{[K]}(\Psi) \in \mathbb{R}^{d\times d},\end{aligned} \quad (7)$$

has basis matrix $A = (a_1, \ldots, a_r) \in \mathbb{R}^{d\times r}$ with $a_j = \mathrm{vec}(\otimes_{k=1}^{K} a_{jk})$ and correlation measures

$$\vartheta = \max\nolimits_{1\leqslant i<j\leqslant r}|a_i^\top a_j|, \quad \delta = \|A^\top A - I_r\|_{\mathrm{S}}, \quad (8)$$

where $d = \prod_{j=1}^{K} d_j$. As $a_i^\top a_j = \prod_{k=1}^{K} a_{ik}^\top a_{jk} = \prod_{k=1}^{K}\sigma_{ij,k}$, the coherence is bounded by $\vartheta \leqslant \prod_{k=1}^{K}\vartheta_k \leqslant \vartheta_{\max}^{K}$. The spectrum norm δ is also bounded by the products of quantities in (6). We summarize these elementary relationships in the following proposition.

Proposition 1: For any set S of tensor modes, define $a_{jS} = \mathrm{vec}(\otimes_{k\in S} a_{jk})$, $A_S = (a_{1S}, \ldots, a_{rS})$, $\vartheta_S = \max_{1\leqslant i<j\leqslant r}|a_{iS}^\top a_{jS}|$ and $\delta_S = \|A_S^\top A_S - I_r\|_{\mathrm{S}}$. Define

$$\mu_S = \max_{j}\min_{k_1,k_2\in S}\max_{i\neq j}\prod_{k\neq k_1, k\neq k_2, k\in S}\sqrt{r}|\sigma_{ij,k}|/\eta_{jk}.$$

as the (leave-two-out) mutual coherence of $\{A_j, j \in S\}$. Then, $\mu_S \in [1, r^{|S|/2-1}]$,

$$\delta_S \leqslant \min_{k\in S}\delta_k, \quad \delta_S \leqslant (r-1)\vartheta_S \leqslant (r-1)\prod_{k\in S}\vartheta_k, \quad (9)$$

$$\delta_S \leqslant \mu_S r^{1-|S|/2}\max\nolimits_{j\leqslant r}\prod_{k\in S}\eta_{jk} \leqslant \mu_S r^{1-|S|/2}\prod_{k\in S}\delta_k. \quad (10)$$

When $S = [K]$, the above inequalities hold with $\{\delta_S, \vartheta_S\}$ replaced by the $\{\delta, \vartheta\}$ in (8).

We note that (10) implies $\delta \leqslant \max_{j\leqslant r}\prod_{k=1}^{K}\eta_{jk} \leqslant \prod_{k=1}^{K}\delta_k$ due to $\mu_* r^{1-K/2} \leqslant 1$. When (most of) the quantities in (6) are small, the products in (10) would be much smaller, so that a_j are nearly orthogonal in (7). This motivates the use of the PCA of (7) to estimate λ_j and a_j,

$$\mathrm{mat}_{[K]}(T) = \textstyle\sum_j \widehat{\lambda}_j^{\mathrm{cpca}}\widehat{u}_j\widehat{u}_j^\top. \quad (11)$$

The following proposition gives explicit justifications of (11) with sharp perturbation bounds.

Proposition 2: Let $d \geqslant r$ and $A \in \mathbb{R}^{d\times r}$ with $\|A^\top A - I_r\|_{\mathrm{S}} \leqslant \delta$. Let $A = \widetilde{U}_1\widetilde{D}_1\widetilde{U}_2^\top$ be the SVD of A, and $U = \widetilde{U}_1\widetilde{U}_2^\top$. Then, $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \leqslant \delta\|\Lambda\|_{\mathrm{S}}$ for all nonnegative-definite matrices Λ in $\mathbb{R}^{r\times r}$.

By Proposition 2, $\widehat{u}_j$ in (11) can be viewed as an estimate of u_j, $1 \leqslant j \leqslant r$, satisfying

$$\|a_j a_j^\top - u_j u_j^\top\|_{\mathrm{S}} = \|A(e_j e_j^\top)A^\top - U(e_j e_j^\top)U^\top\|_{\mathrm{S}} \leqslant \delta. \quad (12)$$

Because $\mathrm{mat}_k(a_j) = a_{jk}\mathrm{vec}(\otimes_{\ell\in[K]\setminus\{k\}} a_{j\ell})^\top$, the natural estimate of a_{jk} based on the $\widehat{u}_j$ in (11) is

$$\widehat{a}_{jk}^{\mathrm{cpca}} = \text{the top left singular vector of } \mathrm{mat}_k(\widehat{u}_j). \quad (13)$$

The following proposition explicitly justifies (13) with a sharp perturbation bound.

Proposition 3: Let $M \in \mathbb{R}^{d_1\times d_2}$ be a matrix with $\|M\|_{\mathrm{F}} = 1$ and a and b be unit vectors respectively in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$. Let $\widehat{a}$ be the top left singular vector of M. Then,

$$\begin{aligned}&\big(\|\widehat{a}\widehat{a}^\top - aa^\top\|_{\mathrm{S}}^2\big) \wedge (1/2) \\ &\leqslant \|\mathrm{vec}(M)\mathrm{vec}(M)^\top - \mathrm{vec}(ab^\top)\mathrm{vec}(ab^\top)^\top\|_{\mathrm{S}}^2.\end{aligned} \quad (14)$$

Proposition 3 is sharp in the sense that equality is attainable in (14) when the right-hand side is less than $1/2$, and that for any $c \in [1/2, 1]$ the maximal distance $\|\widehat{a}\widehat{a}^\top - aa^\top\|_{\mathrm{S}} = 1$ is attainable for some $\{M, a, b\}$ with the right-hand side of (14) being exactly c.

The CPCA is the two-step procedure given by the PCA in (11) and SVD in (13). With $\mathrm{vec}(M) = \widehat{u}_j$, $a = a_{jk}$ and $b = \otimes_{\ell\neq k} a_{j\ell}$, Proposition 3 asserts that the second step of the CPCA is a contraction once the first step yields an estimate of $a_j = \otimes_k a_{jk}$ within 45 degrees. As the correlations among a_j is much smaller than those among a_{jk} in each mode, this condition is much more explicit and of a much weaker form than those in the literature for the estimation of a_{jk} after random projection [29], [48]. By the perturbation bounds in Propositions 2 and 3 and Wedin's perturbation theorem, in the noiseless case ($\sigma = 0$ and $n \to \infty$) with $\Psi = 0$ in (2)

$$\begin{aligned}&|\widehat{\lambda}_j^{\mathrm{cpca}} - \lambda_j| \leqslant \delta\lambda_1, \\ &\big(\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}^2\big) \wedge (1/2) \leqslant (1 + 2\lambda_1/\lambda_{j,\pm})^2\delta^2,\end{aligned} \quad (15)$$

so that the CPCA takes advantage of the multiplicative coherence of $\otimes_{k=1}^{K} a_{jk}$ in view of the product bounds for δ in Proposition 1. We state the CPCA as Algorithm 1 as follows.

Algorithm 1 Composite PCA (CPCA) for Pairwise Symmetric Tensors

Input: noisy tensor $T = n^{-1}\sum_{i=1}^{n}\mathcal{X}_i \otimes \mathcal{X}_i$, CP rank r

1: Formulate T to be a $d \times d$ matrix $\mathrm{mat}_{[K]}(T)$ as in (7) with $d = \prod_{k=1}^{K} d_k$

2: Compute the r top $\widehat{\lambda}_j^{\mathrm{cpca}}$ and $\widehat{u}_j$ in the eigenvalue decomposition of $\mathrm{mat}_{[K]}(T)$ as in (11)

3: Compute $\widehat{a}_{jk}^{\mathrm{cpca}}$ as the top left singular vector of $\mathrm{mat}_k(\widehat{u}_j) \in \mathbb{R}^{d_k\times(d/d_k)}$ as in (13)

Output: $\widehat{a}_{jk}^{\mathrm{cpca}}$, $\widehat{\lambda}_j^{\mathrm{cpca}}$, $j = 1, \ldots, r$, $k = 1, \ldots, K$

After obtaining a warm start through the CPCA (Algorithm 1), we propose to use the ICO (Algorithm 2 below) to refine the solution. The ICO can be viewed as an extension of HOOI [37], [39] and the iterative projection algorithm in [36] to undercomplete ($r < d_{\min}$) and non-orthogonal CP decompositions. However, ICO differentiates from these methods and the alternating least squares in the following important way: In updating the model-k basis vector a_{jk}, the ICO projects the observed tensor T to the orthogonal complements of the span of $\{a_{i\ell}, i \neq j, i \leqslant r\}$ in $\mathbb{R}^{d_\ell}$ for all $\ell \neq k$ simultaneously from $2(K-1)$ sides. Here the word "concurrent" in ICO refers to the feature that the projections take place in all modes $\ell \neq k$ at the same time point/step in the computational iterations. In contexts where "time" has special meaning such as time series, iterative simultaneous orthogonalization can be used instead of ICO. Given estimates $\widetilde{A}_\ell = (\widetilde{a}_{1\ell}, \ldots, \widetilde{a}_{r\ell})$ for the mode-ℓ basis matrix $A_\ell = (a_{1\ell}, \ldots, a_{r\ell}) \in \mathbb{R}^{d_\ell\times r}$, this is done by projecting T to the nonnegative-definite

$$\widetilde{T}_{jk} = T \times_{l\in[2K]\setminus\{k,K+k\}} \widetilde{b}_{jl}^\top \approx \lambda_j a_{jk}a_{jk}^\top + \widetilde{\Psi}_{jk} \quad (16)$$

with $\widetilde{B}_\ell = (\widetilde{b}_{1\ell}, \ldots, \widetilde{b}_{r\ell}) = \widetilde{A}_\ell(\widetilde{A}_\ell^\top\widetilde{A}_\ell)^{-1}$ and $\widetilde{b}_{j,K+\ell} = \widetilde{b}_{jl}$, as $A_\ell^\top\widetilde{B}_\ell \approx I_r$ when $\widetilde{A}_\ell \approx A_\ell$. Thus, it is natural to update

a_{jk} using the top eigenvector of $\widetilde{T}_{jk}$. This is the ICO in Algorithm 2 below.

Algorithm 2 Iterative Concurrent Orthogonalization (ICO) for Pairwise Symmetric Tensors

---

Input: noisy tensor $T = n^{-1}\sum_{i=1}^{n}\mathcal{X}_i \otimes \mathcal{X}_i$, CP rank r, warm-start $\widehat{a}_{jk}^{(0)} \in \mathbb{R}^{d_k}, j \in [r], k \in [K]$, tolerance parameter $\epsilon > 0$, maximum number of iterations M

Compute $(\widehat{b}_{1k}^{(1)}, \ldots, \widehat{b}_{rk}^{(1)}) \in \mathbb{R}^{d_k \times r}$ as the right inverse of $(\widehat{a}_{1k}^{(0)}, \ldots, \widehat{a}_{rk}^{(0)})^\top, k \in [K]$; Set $m = 0$

repeat

  Set $m = m + 1$

  for $k = 1$ to K

    for $j = 1$ to r

      Compute $T_{jk}^{(m)} = T \times_{l \in [2K]\setminus\{k, K+k\}} (\widehat{b}_{jl}^{(m)})^\top$ in $\mathbb{R}^{d_k \times d_k}$ as in (16), $b_{j,K+l}^{(m)} = b_{jl}^{(m)}$

      Compute $\widehat{a}_{jk}^{(m)}$ as the top eigenvector of $T_{jk}^{(m)}$

    end for

    Compute $(\widehat{b}_{1k}^{(m)}, \ldots, \widehat{b}_{rk}^{(m)})$ as the right inverse of $(\widehat{a}_{1k}^{(m)}, \ldots, \widehat{a}_{rk}^{(m)})^\top$

    Set $(\widehat{b}_{1k}^{(m+1)}, \ldots, \widehat{b}_{rk}^{(m+1)}) = (\widehat{b}_{1k}^{(m)}, \ldots, \widehat{b}_{rk}^{(m)})$

  end for

until $\max_{j,k} \|\widehat{a}_{jk}^{(m)}\widehat{a}_{jk}^{(m)\top} - \widehat{a}_{jk}^{(m-1)}\widehat{a}_{jk}^{(m-1)\top}\|_{\mathrm{S}} \leqslant \epsilon$ or $m = M$

Output: $\widehat{a}_{jk}^{\mathrm{ico}} = \widehat{a}_{jk}^{(m)}$, $\widehat{\lambda}_j^{\mathrm{ico}} = T \times_{k=1}^{2K} (\widehat{b}_{jk}^{(m)})^\top$, $j = 1, \ldots, r,\ k = 1, \ldots, K$

Proposition 4: Let $T^* = \mathbb{E}[T]$ with the tensor T in (2). Given $\widetilde{A}_\ell = (\widetilde{a}_{1\ell}, \ldots, \widetilde{a}_{r\ell}), \ell \in [K]\setminus\{k\}$, let $\widetilde{a}_{jk}^*$ be the top eigenvector of $\widetilde{T}_{jk}^* = T^* \times_{l \in [2K]\setminus\{k, K+k\}} \widetilde{b}_{jl}^\top \in \mathbb{R}^{d_k \times d_k}$ with the $\widetilde{b}_{jl}$ in (16). Then,

$$\begin{aligned} &\|a_{jk}a_{jk}^\top - \widetilde{a}_{jk}^*\widetilde{a}_{jk}^{*\top}\|_{\mathrm{S}} \\ \leqslant &2\big(1+\delta_k\big)(\lambda_1/\lambda_j) \prod_{\ell \in [K]\setminus\{k\}} (\widetilde{\phi}_\ell/(1-\widetilde{\phi}_\ell)_+)^2, \end{aligned}$$

where $\widetilde{\phi}_\ell = \widetilde{\psi}_\ell / \big(\sqrt{(1-\delta_\ell)(1-1/(4r))} - \sqrt{r}\widetilde{\psi}_\ell\big)_+$ with $\widetilde{\psi}_\ell = \max_{j \leqslant r} \|\widetilde{a}_{j\ell}\widetilde{a}_{j\ell}^\top - a_{j\ell}a_{j\ell}^\top\|_{\mathrm{S}}$.

The perturbation bound in Proposition 4 explicitly proves the power of concurrent orthogonalization: In terms of the angle between the one-dimensional spaces generated by $\widehat{a}_{jk}^{\mathrm{ico}}$ and a_{jk} and up to some scaling constants, the error in the estimation of a_{jk} in each step is bounded by a product of $2(K-1)$ carryover errors in all other modes in the noiseless case $\Psi = 0$ in model (2), i.e. $\sigma = 0$ and $n \to \infty$ in model (1). In this sense, the ICO error propagates in the order of $2(K-1) > 1$, which implies high order contraction. See Subsection II-C for a more detailed discussion in a comparison between the ICO and the alternating least squares in closely related model (4). As in the analysis of accelerated gradient descent in which the error propagates in the second order, the ICO is expected to achieve ϵ accuracy within $\log\log(1/\epsilon)$ iterations in the noiseless case in model (2). This property of the ICO is confirmed in Theorem 1 and extended to the noisy case in Theorem 3 below in Section III.

B. General High Order Tensors

In Section II-A, we focus on $2K$-th order tensors which can be unfolded as a symmetric matrix. In this section, we extend the CPCA and ICO algorithms to general N-th order tensors.

In model (4), we present the following proposition as an extension of Proposition 2. It also covers the study of the CPCA of spiked covariance tensors developed in Section II-A. Similar to Section II-A, Proposition 1, Proposition 5 and Proposition 3 together provide heuristic justifications for the CPCA in Algorithm 3 below and a road map to study it in model (4).

Proposition 5: Let $A \in \mathbb{R}^{d_1 \times r}$ and $B \in \mathbb{R}^{d_2 \times r}$ with $\|A^\top A - I_r\|_{\mathrm{S}} \vee \|B^\top B - I_r\|_{\mathrm{S}} \leqslant \delta$ and $d_1 \wedge d_2 \geqslant r$. Let $A = \widetilde{U}_1\widetilde{D}_1\widetilde{U}_2^\top$ be the SVD of A, $U = \widetilde{U}_1\widetilde{U}_2^\top$, $B = \widetilde{V}_1\widetilde{D}_2\widetilde{V}_2^\top$ the SVD of B, and $V = \widetilde{V}_1\widetilde{V}_2^\top$. Then, $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \leqslant \delta\|\Lambda\|_{\mathrm{S}}$ for all nonnegative-definite matrices Λ in $\mathbb{R}^{r\times r}$, and $\|AQB^\top - UQV^\top\|_{\mathrm{S}} \leqslant \sqrt{2}\delta\|Q\|_{\mathrm{S}}$ for all $r \times r$ matrices Q.

We note that in Proposition 5, U is a function of A and V is the same function of B.

Algorithm 3 Composite PCA (CPCA) for General N-th Order Tensors

---

Input: noisy tensor $T = \sum_{j=1}^{r}\lambda_j \otimes_{k=1}^{N} a_{jk} + \Psi \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, CP rank r, $S \subset [N]$

1: If $S = \varnothing$, pick S to maximize $\min(d_S, d/d_S)$ with $d_S = \prod_{k \in S} d_k$ and $d = \prod_{k=1}^{N} d_k$

2: Unfold T to be a $d_S \times (d/d_S)$ matrix $\mathrm{mat}_S(T)$

3: Compute $\widehat{\lambda}_j^{\mathrm{cpca}}, \widehat{u}_j, \widehat{v}_j$ as the top components in the SVD $\mathrm{mat}_S(T) = \sum_j \widehat{\lambda}_j^{\mathrm{cpca}}\widehat{u}_j\widehat{v}_j^\top$

4: Compute $\widehat{a}_{jk}^{\mathrm{cpca}}$ as the top left singular vector of $\mathrm{mat}_k(\widehat{u}_j)$, $k \in S$, or $\mathrm{mat}_k(\widehat{v}_j)$, $k \in S^c$

Output: $\widehat{a}_{jk}^{\mathrm{cpca}},\ \widehat{\lambda}_j^{\mathrm{cpca}},\ j = 1, \ldots, r,\ k = 1, \ldots, N$

In practice, a sensible way to unfold T is to form a matrix as square as possible with input $S = \varnothing$ in Algorithm 3. As in Algorithm 2 we propose to use $\widehat{a}_{jk}^{(0)} = \widehat{a}_{jk}^{\mathrm{cpca}}$ as warm-start of the ICO in Algorithm 4 below.

Remark 1: For order 3 tensors, either $|S| = 1$ or $|S^c| = 1$ in Step 1 of Algorithm 3. Assume $d_1 \geqslant d_2 \vee d_3$ for definiteness so that we choose $S = \{1\}$. The CPCA exhibits advantage in terms of coherence by Proposition 1 if and only if $\delta_1 < \delta_2 \vee \delta_3$, e.g when $\delta_k \downarrow d_k$.

Similar to Proposition 4, we present a fresh Proposition 6 to describe the high order of error propagation in the ICO iterations in model (4).

Proposition 6: Let $T^* = \mathbb{E}[T]$ with the tensor T in (4). Given $\widetilde{A}_\ell = (\widetilde{a}_{1\ell}, \ldots, \widetilde{a}_{r\ell}), \ell \in [N]\setminus\{k\}$, let $(\widetilde{b}_{1\ell}, \ldots, \widetilde{b}_{r\ell}) = \widetilde{A}_\ell(\widetilde{A}_\ell^\top\widetilde{A}_\ell)^{-1}$, $\widetilde{T}_{jk}^* = T^* \times_{l \in [N]\setminus\{k\}} \widetilde{b}_{jl}^\top \in \mathbb{R}^{d_k}$, $\widetilde{a}_{jk}^* = \widetilde{T}_{jk}^*/\|\widetilde{T}_{jk}^*\|_2$ and $\widetilde{\lambda}_j^* = T^* \times_{l \in [N]} \widetilde{b}_{jl}^\top$. Then,

$$2 - 2\big|a_{jk}^\top\widetilde{a}_{jk}^*\big| \leqslant 2(r-1)\big(1+\delta_k\big)\Bigg(\frac{\lambda_1}{\lambda_j}\prod_{\ell \in [N]\setminus\{k\}}\frac{\widetilde{\phi}_\ell}{1-\widetilde{\phi}_\ell}\Bigg)^2,$$

$$\big|\widetilde{\lambda}_j^*/\lambda_j - 1\big| \leqslant \sum_{\ell=1}^{N}\widetilde{\phi}_\ell + (r-1)(\lambda_1/\lambda_j)\prod_{\ell=1}^{N}\widetilde{\phi}_\ell,$$

Algorithm 4 Iterative Concurrent Orthogonalization (ICO) for General N-th Order Tensors

Input: noisy tensor $T=\sum_{j=1}^{r}\lambda_j\otimes_{k=1}^{N}a_{jk}+\Psi\in\mathbb{R}^{d_1\times\cdots\times d_N}$, CP rank r, warm-start $\widehat{a}_{jk}^{(0)}, j\in[r], k\in[N]$, tolerance parameter $\epsilon>0$, maximum number of iterations M

1: Compute $(\widehat{b}_{1k}^{(1)},\ldots,\widehat{b}_{rk}^{(1)})$ as the right inverse of $(\widehat{a}_{1k}^{(0)},\ldots,\widehat{a}_{rk}^{(0)})^\top, k\in[N]$; Set $m=0$

2: **repeat**

3: Set $m=m+1$

4: **for** $k=1$ to N

5: **for** $j=1$ to r

6: Compute $T_{jk}^{(m)}=T\times_{l\in[N]\setminus\{k\}}(\widehat{b}_{jl}^{(m)})^\top\in\mathbb{R}^{d_k}$

7: Compute $\widehat{a}_{jk}^{(m)}=T_{jk}^{(m)}/\|T_{jk}^{(m)}\|_2$

8: **end for**

9: Compute $(\widehat{b}_{1k}^{(m)},\ldots,\widehat{b}_{rk}^{(m)})$ as the right inverse of $(\widehat{a}_{1k}^{(m)},\ldots,\widehat{a}_{rk}^{(m)})^\top$

10: Set $(\widehat{b}_{1k}^{(m+1)},\ldots,\widehat{b}_{rk}^{(m+1)})=(\widehat{b}_{1k}^{(m)},\ldots,\widehat{b}_{rk}^{(m)})$

11: **end for**

12: **until** $\max_{j,k}\|\widehat{a}_{jk}^{(m)}\widehat{a}_{jk}^{(m)\top}-\widehat{a}_{jk}^{(m-1)}\widehat{a}_{jk}^{(m-1)\top}\|_{\mathrm{S}}\leqslant\epsilon$ or $m=M$

Output: $\widehat{a}_{jk}^{\mathrm{ico}}=\widehat{a}_{jk}^{(m)}$, $\widehat{\lambda}_j^{\mathrm{ico}}=\left|T\times_{k=1}^{N}(\widehat{b}_{jk}^{(m)})^\top\right|$, $j\in[r], k\in[N]$

where $\widetilde{\phi}_\ell=\widetilde{\psi}_\ell/\big(\sqrt{1-\delta_\ell}-\sqrt{r}\widetilde{\psi}_\ell\big)_+$ with $\widetilde{\psi}_\ell=\max_{j\leqslant r}\big(2-2\big|\widetilde{a}_{j\ell}^\top a_{j\ell}\big|\big)^{1/2}$.

We note that $2-2|a^\top b|=\min_\pm\|a\pm b\|_2^2$ and that each a_{jk} is identifiable only up to a $\pm$ sign. Again, by Proposition 6, the ICO is expected to have a super-linear computational convergence under the loss (5) in the noiseless case $\Psi=0$ in model (4). This is confirmed in Theorem 4 and extended to the noisy case in Theorem 6 below in Section III.

C. Error Propagation in ICO and Alternating LSE

The merit of the ICO can be more directly seen from a comparison with alternating least squares in model (4), $T=T^*+\Psi$ with target tensor $T^*=\sum_{j=1}^{r}\lambda_j\otimes_{k=1}^{N}a_{jk}$ and noise Ψ. Given estimates $\widehat{a}_{jk}^{(m)}, k>1$, the LSE of $A_1\Lambda=(\lambda_1a_{11},\ldots,\lambda_ra_{r1})$ is

$$\begin{aligned}&\mathrm{mat}_1(T)\widehat{B}_{-1}^{\mathsf{LS}}\\&=\arg\min_{M\in\mathbb{R}^{d_1\times r}}\left\|\mathrm{mat}_1(T)-M\big(\widehat{A}_{-1}^{(m)}\big)^\top\right\|_{\mathrm{HS}}^2\\&=A_1\Lambda+A_1\Lambda\big(A_{-1}-\widehat{A}_{-1}^{(m)}\big)\widehat{B}_{-1}^{\mathsf{LS}}+\mathrm{mat}_1(\Psi)\widehat{B}_{-1}^{\mathsf{LS}},\end{aligned}$$

where $\widehat{B}_{-1}^{\mathsf{LS}}$ is the right inverse of $\widehat{A}_{-1}^{(m)}=\big(\widehat{a}_{1,-1}^{(m)},\ldots,\widehat{a}_{r,-1}^{(m)}\big)\in R^{d_{-1}\times r}$ with $\widehat{a}_{j,-1}^{(m)}=\mathrm{vec}\big(\otimes_{k=2}^{N}\widehat{a}_{jk}^{(m)}\big)$ and $d_{-1}=d/d_1$. Because $\widehat{A}_{-1}^{(m)}-A_{-1}$ is an $(N-1)$-degree polynomial of the carryover errors $\widehat{a}_{jk}^{(m)}-a_{jk}$ and the polynomial has a nonvanishing linear term, the leading term of the bias $\mathrm{mat}_1(T^*)\widehat{B}_{-1}^{\mathsf{LS}}-A_1\Lambda$ of the LSE is linear in the carryover error. In comparison, in the ICO, the right inverse is taken in Algorithm 4 in individual modes before tensor multiplication,

$$\begin{aligned}\widehat{A}_1^{(m+1)}\widehat{\Lambda}^{(m+1,1)}&=\mathrm{mat}_1(T)\widehat{B}_{-1}^{(m)}\\&=\mathrm{mat}_1(T^*)\widehat{B}_{-1}^{(m)}+\mathrm{mat}_1(\Psi)\widehat{B}_{-1}^{(m)},\end{aligned}$$

where $\widehat{B}_{-1}^{(m)}=\big(\widehat{b}_{1,-1}^{(m)},\ldots,\widehat{b}_{r,-1}^{(m)}\big)\in R^{d_{-1}\times r}$ with $\widehat{b}_{j,-1}^{(m)}=\mathrm{vec}\big(\otimes_{k=2}^{N}\widehat{b}_{jk}^{(m)}\big)$, $\widehat{\Lambda}^{(m+1,1)}$ is a diagonal matrix to normalize the estimated basis vectors to $\big\|\widehat{a}_{j1}^{(m+1)}\big\|_2=1$. The noise terms $\mathrm{mat}_1(\Psi)\widehat{B}_{-1}^{\mathsf{LS}}$ and $\mathrm{mat}_1(\Psi)\widehat{B}_{-1}^{(m)}$ are comparable between the two methods. However, as

$$\begin{aligned}&\mathrm{mat}_1(T^*)\widehat{B}_{-1}^{(m)}\\&=\left(a_{11}\lambda_1\prod_{k=2}^{N}a_{1k}^\top\widehat{b}_{1k}^{(m)},\ldots,a_{r1}\lambda_r\prod_{k=2}^{N}a_{rk}^\top\widehat{b}_{rk}^{(m)}\right)\\&\quad+\left(\sum_{j=2}^{r}a_{j1}\lambda_j\prod_{k=2}^{N}a_{jk}^\top\widehat{b}_{1k}^{(m)},\ldots,\sum_{j=1}^{r-1}a_{j1}\lambda_j\prod_{k=2}^{N}a_{jk}^\top\widehat{b}_{rk}^{(m)}\right)\end{aligned}$$

with $a_{j_1k}^\top b_{j_2k}^{(m)}=I\{j_1=j_2\}+(a_{j_1k}-a_{j_1k}^{(m)})^\top b_{j_2k}^{(m)}$, the leading term in the bias of the ICO, as the second term above, is a homogeneous polynomial of degree $N-1$ in terms of the carryover error. We note that the errors in the diagonal $\prod_{k=2}^{N}a_{jk}^\top\widehat{b}_{jk}^{(m)}$ is linear in terms of the carryover error but they are absorbed into $\widehat{\Lambda}^{(m+1,1)}$.

In summary, the alternating least squares operator $\widehat{B}_{-1}^{\mathsf{LS}}$ is the inverse of tensor product, while the ICO operator $\widehat{B}_{-1}^{(m)}$ is the tensor product of inverses in $N-1$ individual modes. Consequently, the bias of an alternating least squares step is proportional to the norm of the carryover error and the bias of an ICO step is proportional to the $(N-1)$-th power of the norm of the carryover error. Meanwhile, the noise terms of the two methods are comparable.

D. Algorithm Complexity

Assume the input tensor is T. Algorithm 1 (CPCA) costs $O(d^2r)$ floating-point operations (flops) for r-truncated eigen decomposition of $\mathrm{mat}_{[K]}(T)$ and $O(d)$ flops for 1-truncated SVD of $\mathrm{mat}_k(\widehat{u}_i)$, so that the total cost of CPCA is $O(d^2r)$. In each iteration of Algorithm 2 (ICO), the calculation of $\widehat{B}_k$ costs $O(d_kr^2)$ flops, the matrix manipulation in step 6 costs $O(d^2)$ flops, and the 1-truncated eigen decomposition of $T_{jk}^{(m)}$ in step 7 costs $O(d_k^2)$ flops. Hence, the total cost per iteration in Algorithm 2 is also $O(d^2r)$. Similarly, in Section II-B, the total cost of Algorithm 3 is $O(r\prod_{k=1}^{N}d_k)$, and the cost of each iteration in Algorithm 4 is also $O(r\prod_{k=1}^{N}d_k)$. In summary, the cost of CPCA and each iteration of ICO is of the order of the product of the CP rank and the number of entries in tensor T.

In a spiked covariance tensor model (1), the top r eigenvalue decomposition of the unfolded covariance tensor $\mathrm{mat}_{[K]}(T)$ is equivalent to the top r SVD of the $n^{1/2}$-normalized unfolded data matrix $(\mathrm{vec}(\mathcal{X}_1),\ldots,\mathrm{vec}(\mathcal{X}_n))/\sqrt{n}\in\mathbb{R}^{d\times n}$. In this sense, Algorithms 1 and 2 can be modified accordingly to adopt matrix SVD. The total cost of the first SVD in Algorithm 1 becomes $O(dnr)$, so that the total cost of Algorithm 1 is $O(dnr)$. Similarly, the total cost per iteration in Algorithm 2 is $O(dnr)$. As the cost to construct covariance tensor T is $O(d^2n)$, it can be computationally more efficient to perform the SVD directly.

While the topic is beyond the scope of this paper, we note that random projection and other remedies can be used to reduce the cost of computing low-rank PCA and SVD when the signal to noise ratio is high.

E. Identification and Estimation of CP Component Groups

In principle, the top r singular space $(\widehat{u}_1, \ldots, \widehat{u}_r)$ in CPCA (Algorithms 1 and 3) might not be uniquely determined; for example, this occurs in the presence of ties in λ_j. In such cases, CPCA and ICO may still be used to identify and estimate CP component groups with tied singular values. To avoid redundancy, we describe the procedures below only for the symmetric tensors in (2).

Suppose there are g groups of singular values with distinct representative values $\lambda_{(1)} > \cdots > \lambda_{(g)} > 0$ and respective group sizes $r_1, \ldots, r_g$, $r_1 + \cdots + r_g = r$. Suppose (2) can be written as

$$T = \textstyle\sum_{i=1}^{g} T_{(i)} + \Psi,$$
$$T_{(i)} = \textstyle\sum_{j \in G_i} \lambda_j \otimes_{k=1}^{2K} a_{jk} \approx \lambda_{(i)} \textstyle\sum_{j \in G_i} \otimes_{k=1}^{2K} u_{jk},$$

where $\{G_1, \ldots, G_g\}$ is a partition of $[r]$ with $|G_i| = r_i$. By Proposition 2, it is reasonable to consider the case where $\{u_{jk}, j \in G_i, i = 1, \ldots, g\}$ are orthonormal for each $k \in [K]$ and

$$\max_{j \leqslant r} \|a_{jk} a_{jk}^\top - u_{jk} u_{jk}^\top\|_{\mathrm{S}} \leqslant \delta^*,$$
$$\|A \Lambda A^\top - \textstyle\sum_{i=1}^{g} \lambda_{(i)} \textstyle\sum_{j \in G_i} u_j u_j^\top\|_{\mathrm{S}} \leqslant \lambda_{(1)} \delta^*$$

with $u_j = \mathrm{vec}\big(\otimes_{k=1}^{K} u_{jk}\big)$ and a certain $\delta^* \approx \delta$. Suppose further that $2\delta^* \lambda_{(1)} < \min_{i \leqslant g}(\lambda_{(i)} - \lambda_{(i+1)})$ with $\lambda_{(g+1)} = 0$. It would then be reasonable to consider clustering of the outputs of CPCA and ICO to identify the groups G_i, e.g. using $\widehat{\lambda}_j^{\mathrm{cpca}}$ by Proposition 2. Given G_j, a group ICO could be used to estimate the individual $T_{(i)}$ and then a Tucker decomposition would give the column space of $A_{G_i k} = (a_{jk}, j \in G_i)$ for each (i, k).

Once a good estimate of the group tensor $T_{(i)}$ becomes available, the identification of individual components $\lambda_j \otimes_{k=1}^{2K} a_{jk}, j \in G_i$, in the group would be feasible if a rank-one component can be identified in the linear span of the group components. This feasibility can be seen from Proposition 7 below. Since the identifiability issue does not require paired CP bases as in (2), Proposition 7 is stated under model (4). Kruskal's Theorem [51] also provides the uniqueness of tensor CP decomposition.

Proposition 7: Let SP $:= \mathrm{span}\{a_1, \ldots, a_r\}$, where $a_j = \mathrm{vec}(\otimes_{k=1}^{N} a_{jk}) \in \mathbb{R}^d$. The elements of SP can be viewed as either length d vectors or $d_1 \times \cdots \times d_N$ tensors. Suppose $N > 2$ and $\delta_k < 1$ for every $k = 1, \ldots, N$ in (6), then every rank-1 tensor in SP is one of a_j's up to a scalar.

The above discussions, written in response to an interesting question raised by referees, seem to deserve further investigation. However, a more comprehensive discussion or further development in this direction is beyond the scope of this paper.

III. Theoretical Properties

A. Spiked Covariance Tensor Models

In this section, we investigate theoretical guarantees of the proposed algorithms for the estimation of the CP basis vectors a_{jk} for the spiked covariance tensor (2) with data in (1). As in (5) we use $\|\widehat{a}_{jk}\widehat{a}_{jk}^\top - a_{jk}a_{jk}^\top\|_{\mathrm{S}} = (1 - (\widehat{a}_{jk}^\top a_{jk})^2)^{1/2} = \sup_{z \perp a_{jk}, \|z\|_2 = 1} |z^\top \widehat{a}_{jk}|$ to measure the distance between $\widehat{a}_{jk}$ and a_{jk}.

We do not impose the orthogonality condition on the mode-k CP basis vectors $\{a_{jk}, j \leqslant r\}$ or even global incoherence condition on $\vartheta_{\max} := \max_k \max_{1 \leqslant i < j \leqslant r} |a_{ik}^\top a_{jk}|$ as in the literature [13], [15], [29], [47], [48]. However, we require the vectorized basis tensors $a_j = \mathrm{vec}(\otimes_{k=1}^{K} a_{jk})$ to satisfy the isometry condition $\delta = \|A^\top A - I_r\|_{\mathrm{S}} < 1$, $A = (a_1, \ldots, a_r)$, or more conveniently the incoherence condition $\vartheta = \max_{i \neq j} |a_i^\top a_j| < 1/r$. We recall that by Proposition 1, δ and ϑ are bounded by the respective products of their mode-k counterparts defined in (6), so that we impose much weaker conditions compared with the existing ones on $\vartheta_{\max}$. In fact, the higher the tensor order K, the faster the convergence rate we offer given $\{r, \delta, \vartheta\}$, and the smaller δ and θ given r and $\vartheta_{\max}$. Our analysis is based on the perturbation bounds in Propositions 2, 3 and 4 in Section II and proper concentration inequalities. For simplicity, we assume $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ with $\lambda_j = w_j^2$ in (1) and (2).

Theorem 1: Suppose Algorithm 1 (CPCA) is applied to the noiseless $T^* = \sum_{j=1}^{r} \lambda_j \otimes_{k=1}^{2K} a_{jk}$ with $a_{j,K+k} = a_{jk}$. Then, (15) holds for the resulting $\widehat{\lambda}_j^{\mathrm{cpca}}$ and $\widehat{a}_{jk}^{\mathrm{cpca}}$. Let $\lambda_{\min,\pm} = \min_{1 \leqslant j \leqslant r} \lambda_{j,\pm}$ be the minimum eigengap. Suppose further that

$$\begin{gathered} 2\max\big\{\delta_{\max}, (\sqrt{r}+1)\psi_0\big\} \leqslant 1, \\ 3(\lambda_1/\lambda_r)\psi_0^{2K-3} \leqslant \rho < 1, \end{gathered} \tag{17}$$

where $\delta_{\max} = \max_{k \leqslant K} \delta_k$ with the δ_k in (6) and $\psi_0 = (1 + 2\lambda_1/\lambda_{\min,\pm})\delta$ with the δ in (8). Let $\gamma_K \in (3 - 3/K, 3)$ be the solution of $\gamma_K^K - 3\gamma_K^{K-1} + 2 = 0$, e.g. $\gamma_3 = 2.732$, $\gamma_4 = 2.919$. If the resulting $\widehat{a}_{jk}^{\mathrm{cpca}}$ are used as the initialization of Algorithm 2 (ICO) with the same data T^*, then

$$\max_{j \leqslant r} \big\|\widehat{a}_{jk}^{(m)} \widehat{a}_{jk}^{(m)\top} - a_{jk}a_{jk}^\top\big\|_{\mathrm{S}} \leqslant \psi_{m,k} = \psi_0 \rho^{\gamma_K^{m \;\, 1\, K \;\, k \;\, 1}}$$

and $\max_{1 \leqslant k \leqslant K} \psi_{m,k} \leqslant \epsilon$ within m iterations, where $m = \lceil K^{-1}\{1 + (\log\gamma_K)^{-1} \log(\log(\psi_0/\epsilon)/\log(1/\rho))\}\rceil$.

Remark 2 (Condition on the Initial Estimator): The constant factors 2 and 3 in (17) are not sharp. In fact, condition (17) is simplified from the following,

$$\frac{2(1+\delta_{\max})(\lambda_1/\lambda_r)\psi_0^{2K-2}}{\big(\sqrt{(1-\delta_{\max})c_r} - (\sqrt{r}+1)\psi_0\big)_+^{2K-2}} \leqslant \rho\psi_0 < \psi_0, \tag{18}$$

with $c_r = 1 - 1/(4r)$. Condition (18) is slightly sharper and actually used in the proof. Here ψ_0 is an error bound for the initial estimator. The essence of our analysis of the ICO is that under (18), $\psi_m \leqslant C_0 \psi_{m-1}^{2K-2}$ for the error bound $\psi_m = \max_{k \leqslant K} \psi_{m,k}$ in the m-th iteration.

Remark 3 (Incoherence Condition): When the minimum eigenvalue gap satisfies $\lambda_{\min,\pm} \gtrsim \lambda_1/r$, condition (17) asserts that the CPCA needs no stronger incoherence condition than

$\vartheta_{\max} = O(r^{-5/(2K)})$, in view of Proposition 1. In comparison, conditions of stronger form are imposed in the literature; For example the initial estimator in [29] requires the incoherence condition $\vartheta_{\max} \leqslant \mathrm{polylog}(d_{\min})/\sqrt{d_{\min}}$ for 3-way tensors. Compared with the previous work, (17) implies a weaker incoherence condition when $r \lesssim d_{\min}^{(K/5)\wedge 1}$.

Theorem 1 explicitly guarantees the high-order convergence of the ICO algorithm with the CPCA initialization in the noiseless case. To the best of our knowledge, the proposed ICO is the first algorithm known to achieve ϵ-accuracy guarantee within $\log\log(1/\epsilon)$ number of iteration passes in non-orthogonal CP models.

We proceed to present the statistical properties of the proposed estimator in the presence of noise, with input data T in (2). Define

$$\mathrm{SNR} = \frac{\mathbb{E}\big\|\sum_{j=1}^r w_j f_{ij} \otimes_{k=1}^K a_{jk}\big\|_{\mathrm{HS}}^2}{\mathbb{E}\|\mathcal{E}_i\|_{\mathrm{HS}}^2}$$

as the signal-to-noise ratio (SNR) in the covariance tensor CP model (1). As $\lambda_j = w_j^2$ and $\mathbb{E}[f_{ij}^2] = 1$,

$$\mathrm{SNR} = \frac{\mathrm{trace}\big(\mathrm{mat}_k(T^*)\big)}{\sigma^2 d} = \frac{\sum_{j=1}^r \lambda_j}{\sigma^2 d} = \frac{r_{\mathrm{eff}}\lambda_1}{\sigma^2 d} \tag{19}$$

with the signal tensor $T^* = \sum_{j=1}^r \lambda_j \otimes_{k=1}^{2K} a_{jk}$, where $r_{\mathrm{eff}} = \sum_{j=1}^r \lambda_j/\lambda_1$, no greater than the CP rank r, can be viewed as the effective rank of T^*.

Theorem 2: Consider spiked covariance tensor model (2) with data in (1), $\lambda_j = \omega_j^2$ and $\delta = \|A^\top A - I_r\|_{\mathrm{S}}$ as in (8). In an event with probability at least $1 - e^{-t}$, Algorithm 1 (CPCA) gives the following error bound for the estimation of the CP basis vectors a_{jk},

$$\begin{aligned} &\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}} \\ &\leqslant (1 + 2\lambda_1/\lambda_{j,\pm})\delta + C(\lambda_1/\lambda_{j,\pm})\big(R^{(0)} + \sqrt{t/n}\big) \end{aligned} \tag{20}$$

for all $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant K$ and $0 \leqslant t \leqslant d$, where C is a numeric constant, $\lambda_{j,\pm}$ is the j-th eigengap, and $R^{(0)} = \sqrt{(r_{\mathrm{eff}}/n)(1 + 1/\mathrm{SNR})(1 + (r_{\mathrm{eff}}/d)/\mathrm{SNR})} \leqslant \sqrt{(r + \sigma^2 d/\lambda_1)(1 + \sigma^2/\lambda_1)/n}$.

The CPCA error bound (20) consists of two parts. The first part involving δ is induced by the non-orthogonality of the vectors a_{jk}, which can be viewed as bias; The second part comes from a concentration bound for the centered random noise tensor $\Psi - \mathbb{E}[\Psi]$, which can be viewed as stochastic error. When the minimum eigengap satisfies $\lambda_{\min,\pm} \gtrsim \lambda_1/r$, Theorem 2 asserts that the CPCA needs no stronger incoherence condition than $\vartheta_{\max} = O(r^{-2/K})$, in view of Proposition 1. As long as $r \lesssim d_{\min}^{K/4}$, this incoherence condition is weaker than those in the existing literature for tensor denoising in CP models [29]. The error bound (20) is dominated by the bias when $\delta \gtrsim R^{(0)}$, and by the stochastic error when $R^{(0)} \gtrsim \delta$. The stochastic error $R^{(0)}$ can be further divided into two components: the impact of the fluctuation of the signal factor f_{ij} represented by the parametric rate $\sqrt{r_{\mathrm{eff}}/n}$, and the impact of the noise $\mathcal{E}_i$ in (1) represented by $\sqrt{(r_{\mathrm{eff}}/n)/\mathrm{SNR}}$. The noise component dominates the stochastic error iff $\mathrm{SNR} > 1$. Still, the consistency of the CPCA in Theorem 2 requires a SNR condition $\mathrm{SNR} \gtrsim r^3/n$, parallel to the condition $\sqrt{\lambda_r/\sigma^2} \geqslant Cr\sqrt{d/n}$ in the scenario considered in [39].

Next, we consider the theoretical properties of the ICO. We assume below for simplicity that $d_1 \leqslant \cdots \leqslant d_K$. Let

$$R_{jk}^{(\mathrm{ideal})} = \big(\sigma^2/\lambda_j + \sigma/\lambda_j^{1/2}\big)\sqrt{d_k/n}. \tag{21}$$

and for $\phi \geqslant 0$ define

$$R_{jk,\phi}^{(\mathrm{ideal})} = R_{jk}^{(\mathrm{ideal})} + (\phi \wedge 1)\textstyle\sum_{\ell \in [K]\setminus\{k\}} R_{j\ell}^{(\mathrm{ideal})}. \tag{22}$$

For constants $\psi_0 \in (0, 1)$ and $C_0 \geqslant 1$, define

$$\begin{aligned} \alpha &= \sqrt{1 - \delta_{\max}} - (r^{1/2} + 1)\psi_0/\sqrt{1 - 1/(4r)}, \\ \rho &= C_{0,\alpha}(\lambda_1/\lambda_r)\psi_0^{2K-3}, \\ \rho_1 &= C_{0,\alpha}\sqrt{(\lambda_1/\lambda_r)r/n}\,\psi_0^{K-2}, \\ \phi_0 &= C_{0,\alpha}\sqrt{2r/(1 - 1/(4r))}R_{rK,1}^{(\mathrm{ideal})}, \end{aligned} \tag{23}$$

with $\delta_{\max} = \max_{k\in[K]}\delta_k$ and $C_{0,\alpha} = C_0\alpha^{2-2K}$. Let $\mathscr{P}_\pm$ be the class of all $r \times r$ diagonal matrices Π_r with $\Pi_r^2 = I_r$.

Theorem 3: Suppose that with a proper numeric constant C_0 and the quantities defined in (21), (22) and (23),

$$\alpha > 0,\ \rho_1 \leqslant \rho < 1,\ C_{0,\alpha}R_{rK,1}^{(\mathrm{ideal})} \leqslant \psi_0 < 1. \tag{24}$$

Let $\Omega_0 = \{\max_{j,k}\|\widehat{a}_{jk}^{(0)}\widehat{a}_{jk}^{(0)} - a_{jk}a_{jk}^\top\|_{\mathrm{S}} \leqslant \psi_0\}$ for any initial estimates $\widehat{a}_{jk}^{(0)}$. Then, Algorithm 2 (ICO) provides

$$\begin{aligned} &\mathbb{P}\Big\{\max_{j,k}\min_{\Pi_r\in\mathscr{P}}\frac{\|\widehat{A}_k^{\mathrm{ico}}\Pi_r - A_k\|_{\mathrm{F}}}{(4r/3)^{1/2}(\epsilon_{rk} \vee \epsilon)} \leqslant 1\Big\} \\ \geqslant\ &\mathbb{P}\Big\{\max_{j,k}\frac{\|\widehat{a}_{jk}^{\mathrm{ico}}\widehat{a}_{jk}^{\mathrm{ico}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}}{\epsilon_{jk} \vee \epsilon} \leqslant 1\Big\} \\ \geqslant\ &\mathbb{P}\big\{\Omega_0\big\} - mrKe^{-2(d_1\wedge\sqrt{n})} \end{aligned} \tag{25}$$

within $m \geqslant m_\epsilon + 3$ iterations, where $\epsilon_{jk} = C_{0,\alpha}R_{jk,\phi_0}^{(\mathrm{ideal})}$, $m_\epsilon = \lceil\log(\log(\epsilon/\psi_0)/\log\rho)/\log 2\rceil$ for $(\epsilon_{r2} \vee \epsilon_0) \wedge \epsilon_{r3} \leqslant \epsilon < \psi_0$ and $m_\epsilon = \lceil\log(\epsilon/\psi_0)/\log\rho\rceil$ for $\epsilon_{r2} \leqslant \epsilon < \epsilon_0 \wedge \epsilon_{r3}$, with $\epsilon_0 = C_{0,\alpha}r/n$. Moreover, (25) holds within $m_{\epsilon_{r2}} + 4$ iterations for $\epsilon = \epsilon_* \vee \sqrt{\epsilon_*\epsilon_0}$ where $\epsilon_* = C_{0,\alpha}(\lambda_1/\lambda_r)\prod_{k=2}^K \epsilon_{rk}^2$. In particular, if Algorithm 1 (CPCA) is used to initialize Algorithm 2 and ψ_0 is taken as the maximum of the right-hand side of (20), then (25) holds with $\mathbb{P}\{\Omega_0\} \geqslant 1 - e^{-t}$.

In Theorem 3, ϵ_{jk} can be viewed as statistical error and ϵ as computational error. It asserts that by iteratively projecting data (and thus the noise) to the direction $b_{j\ell}$ in mode-ℓ for all $\ell \neq k$, $(b_{1\ell}, \ldots, b_{r\ell}) = A_\ell(A_\ell^\top A_\ell)^{-1}$, Algorithm 2 (ICO) effectively strengthens SNR from (19) to $r\lambda_1/(\sigma^2 d_k)$ in the estimation of a_{jk} while quickly reduces the bias to below the level of stochastic error. As expected from the $\log\log(1/\epsilon)$ convergence in Theorems 1 and 3, the algorithm typically converges within very few steps in our practical implementations.

Theorem 3 indicates that Algorithm 2 converges linearly in its last phase with $\epsilon_{r2} \leqslant \epsilon < \epsilon_0 \wedge \epsilon_{r3}$. However, if we treat the covariance tensor T in model (2) as a general order $2K$ tensor and apply Algorithm 4, high-order convergence can be also achieved in this last phase. The constant $\log 2$ in the definition of m_ϵ is conservative. In fact, by the proof of Theorem 3, Algorithm 2 converges in multiple phases beginning from order $2K - 2$ convergence in its first phase.

The right-hand side of (2) can be improved to $\mathbb{P}\{\Omega_0\} - rKe^{-2(d_1 \wedge \sqrt{n})}$ if the constants in (23) are raised by a factor of at most order K if we apply the probability calculation in the proof of Theorem 6. The Gaussian assumption can be replaced by sub-Gaussian in our analysis.

In Theorem 3, ψ_0 is the required accuracy of the initial estimator. Given $\{C_0, r, \delta_{\max}, \lambda_1/\lambda_r\}$, the first two conditions in (24) hold when ψ_0 is sufficiently small, so that the third condition in (24) is a signal strength condition in terms of $R_{rK,1}^{(\mathrm{ideal})} = \max_{j,\phi} R_{jk,\phi}^{(\mathrm{ideal})}$. In view of the definition of α in (23), condition (24) requires $r^{1/2}\psi_0$ be small, with an extra factor $r^{1/2}$ on the initial error in the estimation of individual basis vectors. This is a technical issue due to the need to invert the estimated $\Sigma_\ell = A_\ell^\top A_\ell$ in our analysis to construct the mode-ℓ projection in the ICO. In practice, if this issue is of concern, one may consider regularized inverse such as by adding a small constant to $\widehat{\Sigma}_\ell$ before computing the inverse or shrinking the singular values of $\widehat{\Sigma}_\ell$ as [29] suggested. If the right-hand side of (20) is taken as ψ_0 for the CPCA initialization, condition (24) can be reduced to an incoherence condition $r^{3/2}\delta \lesssim 1$ when $\lambda_1 \asymp \lambda_r \asymp r\lambda_{j,\pm}$ and σ^2 and $1/n$ are sufficiently small.

When $\sqrt{r}(R_{rK}^{(\mathrm{ideal})})^2 \lesssim R_{jk}^{(\mathrm{ideal})}$, the statistical error $\epsilon_{jk} \lesssim R_{jk}^{(\mathrm{ideal})}$. In the literature of tensor factor models with a Tucker structure [9], [36], the estimation of a_{jk} may achieve faster convergence rate than $O_{\mathbb{P}}(n^{-1/2})$ when $\lambda_j = w_j^2$ is sufficiently large. Similarly, (25) may also converge faster than $O_{\mathbb{P}}(n^{-1/2})$.

Remark 4 (Statistical Optimality): The performance bound in (25) is free of rank r. The rate $R_{jk}^{(\mathrm{ideal})}$ matches the statistical lower bound of [52] and [36] under specific rank one spiked covariance models respectively for matrix and tensor data. Therefore, under proper conditions, the proposed method (Algorithm 2) achieves the minimax optimal rate of convergence in the estimation of a_{jk}.

B. General High Order Tensors

In the noiseless case with $\Psi = 0$ in (4), the extension of Theorem 1 to Algorithms 3 and 4 is straightforward, which explicitly guarantees the high-order convergence of ICO with CPCA initialization. As in Proposition 1 let $a_{jS} = \mathrm{vec}(\otimes_{k\in S} a_{jk})$, $A_S = (a_{1S}, \ldots, a_{rS})$, $\Sigma_S = A_S^\top A_S$ and $\delta_S = \|\Sigma_S - I_r\|_{\mathrm{S}}$ for any nonempty subset S of $[N] = \{1, \ldots, N\}$.

Theorem 4: Suppose Algorithm 3 (CPCA) is applied to the noiseless data $T^* = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk}$ through the SVD of $\mathrm{mat}_S(T^*)$ for some nontrivial subset $S \subset [N]$. Let $\psi_0 = (\sqrt{2} + 4\lambda_1/\lambda_{\min,\pm})\delta$ with $\delta = \delta_S \vee \delta_{S^c}$, where $S^c = [N]\backslash S$. Then,

$$\begin{aligned} |\widehat{\lambda}_j^{\mathrm{cpca}} - \lambda_j| &\leqslant \sqrt{2}\delta\lambda_1, \\ \big(\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}^2\big) \wedge (1/2) &\leqslant \psi_0^2/2, \end{aligned} \tag{26}$$

for the resulting $\widehat{\lambda}_j^{\mathrm{cpca}}$ and $\widehat{a}_{jk}^{\mathrm{cpca}}$. Suppose further that for $\delta_{\max} = \max_{k\leqslant N}\delta_k$,

$$\begin{aligned} &3\max\big\{\delta_{\max}, (\sqrt{r}+1)\psi_0\big\} \leqslant 1, \\ &4\sqrt{r-1}(\lambda_1/\lambda_r)\psi_0^{N-2} \leqslant \rho < 1. \end{aligned} \tag{27}$$

Let $\gamma_N \in (2 - 2/N, 2)$ be the solution of $\gamma_N^N - 2\gamma_N^{N-1} + 1 = 0$, e.g. $\gamma_3 = 1.618$, $\gamma_4 = 1.839$. If the resulting $\widehat{a}_{jk}^{\mathrm{cpca}}$ is used as the initialization of Algorithm 4 (ICO), then

$$\begin{aligned} \max_{j\leqslant r}\big(2 - 2\big|a_{jk}^\top\widehat{a}_{jk}^{(m)}\big|\big)^{1/2} &\leqslant \psi_{m,k} = \psi_0\rho^{\gamma_N^{m \quad 1 \; N \quad k \quad 1}}, \\ \max_{j\leqslant r}\big|\widehat{\lambda}_j^{(m)}/\lambda_j - 1\big| &\leqslant \textstyle\sum_{k=1}^N \psi_{m,k} + \rho\psi_{m,N}, \end{aligned}$$

and $\max_{1\leqslant k\leqslant N}\psi_{m,k} \leqslant \epsilon$ within m iterations, where $m = \lceil N^{-1}\{1 + (\log\gamma_N)^{-1}\log(\log(\psi_0/\epsilon)/\log(1/\rho))\}\rceil$.

Remark 5: Condition (27) specifies the required incoherence condition via δ. Again, the constant factors 3 and 4 in the condition is not sharp, as (27) is simplified from the following condition actually used in the proof,

$$\frac{\sqrt{2(r-1)(1+\delta_{\max})}(\lambda_1/\lambda_r)\psi_0^{N-1}}{\big((1-\delta_{\max})^{1/2} - (\sqrt{r}+1)\psi_0\big)_+^{N-1}} \leqslant \rho\psi_0 < \psi_0. \tag{28}$$

As we have discussed in Remark 2, such conditions guarantee the high-order contraction of the ICO and the resulting $\log\log(1/\epsilon)$ rate.

Now consider statistical properties of Algorithms 3 and 4 for general (asymmetric) tensors $T = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk} + \Psi$ in model (4), where $a_{jk} \in \mathbb{R}^{d_k}$ are basis vectors with $\|a_{jk}\|_2 = 1$, and Ψ is the noise tensor. Similar to the analysis of the spiked covariance tensor model given by (1) and (2), we assume for notational simplicity $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$.

Theorem 5: Let $T = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk} + \Psi$ as in (4). Suppose $\Psi \in \mathbb{R}^{d_1\times\cdots\times d_N}$ has i.i.d $N(0,\sigma^2)$ entries. Then, in an event with probability at least $1 - e^{-2d_S - 2(d/d_S)}$, Algorithm 3 (CPCA) gives the following bound in the estimation of the CP basis vectors a_{jk}, $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant N$,

$$\begin{aligned} &\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}} \\ &\leqslant (1 + 2\sqrt{2}\lambda_1/\lambda_{j,\pm})\delta + 6\sigma(\sqrt{d_S} + \sqrt{d/d_S})/\lambda_{j,\pm} \end{aligned} \tag{29}$$

where $\delta = \|A_S^\top A_S - I\|_{\mathrm{S}} \vee \|A_{S^c}^\top A_{S^c} - I_r\|_{\mathrm{S}}$ as in Theorem 4 and $\lambda_{j,\pm} = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$ are the eigengaps with $\lambda_0 = 2\lambda_1$ and $\lambda_{r+1} = 0$.

The second term in (29), representing the stochastic error, describes the required SNR for the CPCA. It is comparable to the SNR for tensor unfolding method in rank one symmetric case [41], which is proved in [53] to match an optimal computational lower bound under certain conditions. Moreover, the SNR condition here is weaker than the perturbation condition of the initialization in [29] when $\lambda_r/\lambda_{\min,\pm} = o(\sqrt{d_{\max}/\log(r)})$, which is typically satisfied for large d_k.

For simplicity, we assume below $d_1 \leqslant \cdots \leqslant d_N$. Let

$$R_{jk}^{*(\mathrm{ideal})} = \sigma\sqrt{d_k}/\lambda_j. \tag{30}$$

and for $\phi \geqslant 0$ define

$$R_{jk,\phi}^{*(\mathrm{ideal})} = R_{jk}^{*(\mathrm{ideal})} + (\phi \wedge 1)\textstyle\sum_{\ell=1}^N R_{j\ell}^{*(\mathrm{ideal})}. \tag{31}$$

For constants $\psi_0 \in (0,1)$, define

$$\begin{aligned} \alpha_* &= \sqrt{1-\delta_{\max}} - (r^{1/2}+1)\psi_0, \\ \rho^* &= 6\alpha_*^{1-N}\sqrt{r-1}(\lambda_1/\lambda_r)\psi_0^{N-2}, \\ \phi_0^* &= (N-1)\alpha_*^{-1}\sqrt{2r}R_{rN,1}^{*(\mathrm{ideal})}. \end{aligned} \tag{32}$$

Theorem 6: Let data T be as in Theorem 5 and $\Omega_0 = \{\max_{j,k}(2-2|a_{jk}^\top \widehat{a}_{jk}^{(0)}|)^{1/2} \leqslant \psi_0\}$ for any initial estimates $\widehat{a}_{jk}^{(0)}$. Let $\mathscr{P}_\pm$ be as in (25). Suppose

$$\alpha_* > 0,\ \rho^* < 1,\ 6\alpha_*^{1-N} R_{rK,1}^{*(\text{ideal})} \leqslant \psi_0 < 1, \tag{33}$$

with the quantities defined in (30), (31) and (32). Then, in an event with probability at least $\mathbb{P}\{\Omega_0\} - e^{-d_N} - \sum_{k=1}^N e^{-d_k}$, Algorithm 4 (ICO) provides

$$|\widehat{\lambda}_j^{\text{ico}}/\lambda_j - 1| \leqslant \epsilon_{jN}^* \vee \epsilon, \tag{34}$$

$$\|\widehat{a}_{jk}^{\text{ico}} \widehat{a}_{jk}^{\text{ico}\top} - a_{jk}a_{jk}^\top\|_{\text{S}} \leqslant \epsilon_{jk}^* \vee \epsilon, \tag{35}$$

$$\min_{\Pi_r \in \mathscr{P}} \|\widehat{A}_k^{\text{ico}}\Pi_r - A_k\|_{\text{F}} \leqslant r^{1/2}(\epsilon_{rk}^* \vee \epsilon), \tag{36}$$

simultaneously for all $1 \leqslant j \leqslant r$ and $1 \leqslant k \leqslant N$, within $m \geqslant m_\epsilon + 3$ iterations, where $\epsilon_{jk}^* = 6\alpha_*^{N-1} R_{jk,\phi_0}^{*(\text{ideal})}$ and $m_\epsilon = \lceil \log(\log(\epsilon/\psi_0)/\log\rho^*)/\log 2 \rceil$ for $\epsilon_{r2}^* \leqslant \epsilon < \psi_0$. Moreover, (34), (35) and (36) hold in the same event within $m_{\epsilon_{r2}} + 4$ iterations for $\epsilon = 6\alpha_*^{1-N}\sqrt{r-1}(\lambda_1/\lambda_r)\prod_{k=2}^N \epsilon_{rk}^*$. If Algorithm 3 (CPCA) is used as initialization, then $\mathbb{P}\{\Omega_0\} \geqslant 1 - \sum_{k=1}^N e^{-2d_k}$ for $\psi_0 = 6[\lambda_1\delta + \sigma(\sqrt{d_S} + \sqrt{d/d_S})]/\lambda_{\min,\pm}$.

We briefly discuss the conditions and conclusions of Theorem 6 as the details are parallel to the discussions below Theorem 3. In Theorem 6, ϵ_{jk}^* can be viewed as statistical error and ϵ as computational error. When $\sqrt{r}(R_{rN}^{*(\text{ideal})})^2 \lesssim R_{jk}^{*(\text{ideal})}$, the statistical error $\epsilon_{jk}^* \lesssim R_{jk}^{*(\text{ideal})}$ is rate minimax. Condition (33) specifies the required strength of the signal and accuracy of the initialization. It guarantees that the ICO has a high-order error contraction effect in the iteration. Ignoring the perturbation error and assuming $\lambda_1 \asymp \lambda_r$, it can be reduced to an incoherence condition $r^{3/2}\delta \lesssim 1$ when CPCA is used as initialization. In addition, the performance bound in (35) is free of CP rank r and matches the statistical lower bound of [39] for rank one noisy tensor model. It shows the optimality of the convergence rate of the proposed ICO (Algorithm 4).

C. Comparison With Existing Theoretical Results

In this subsection, we compare the proposed Algorithms 3 and 4 with existing theories of tensor decomposition methods. Several important implications are provided, and comparisons in incoherence condition, iteration complexity, and statistical error bounds are summarized in Table I. For simplicity, the following discussion assumes model (4) with $\lambda_1 \asymp \lambda_r$, $\lambda_{\min,\pm} \asymp \lambda_r/r$ and Gaussian noise Ψ.

1) Super-Linear Convergence: In the absence of noise, the proposed algorithm attains ϵ accuracy within $O(\log\log(1/\epsilon))$ iterations. In the noisy setting, the algorithm reaches an ideal statistical accuracy within an iterated logarithmic number of iterations. The perturbation bounds in Propositions 4 and 6 explicitly give the order of convergence for ICO: Up to some scaling constants, the error in the estimation of a_{jk} in each step is bounded by the product of the up-to-date errors in all other modes. As in the analysis of Nesterov's acceleration of gradient descent, this multiplicative nature of error propagation leads to a $\log\log(1/\epsilon)$ convergence rate. In alternating least squares [29] and HOOI, the error propagation is linear due to tensor unfolding so that the convergence rate is of the order $\log(1/\epsilon)$. Still, in certain problems where computationally feasible initialization leads to very high signal-to-noise ratio, one-step least squares or HOOI update would reduce the error to the level of statistical efficiency [36], [39], [55].

2) Statistical Accuracy: While our theoretical analysis is focused on the estimation of individual basis vectors a_{jk}, our results have direct implications on the estimation under different loss functions or of related functions beyond the explicite statements of Theorems 6. For example, for the estimation of the entire tensor $T^* = \mathbb{E}[T]$ in model (4), Theorems 6 directly yields the Frobenius error bound

$$\|\widehat{T} - T^*\|_{\text{F}} \lesssim K\lambda_1 r^{1/2}(\epsilon_{rK}^* \vee \epsilon).$$

Compared with [29], Theorems 6 provide comparable or sharper error bounds under their conditions. The error bound of the CP decomposition algorithms in [29] is $\|\widehat{A}_k\Pi_r - A_k\|_F \leqslant C\sqrt{r}\|\Psi\|_*/\lambda_r$, where $\|\Psi\|_*$ is the tensor spectrum norm, with $\|\Psi\|_* \asymp \sigma\sqrt{d_1 + \cdots + d_N}$ in the Gaussian case. In comparison, Theorem 6 provides $\|\widehat{A}_k^{\text{ico}}\Pi_r - A_k\|_{\text{F}} \leqslant C\sigma\sqrt{d_k r}/\lambda_r$ for ICO with CPCA initialization, matching the statistical lower bound of [39].

3) Incoherence Condition for Initialization: Existing initialization approaches [29], [54], [56] focus on randomized projection in each tensor mode simultaneously to reduce the original data tensor to matrices of effective rank near 1, followed by matrix SVD to obtain rough estimates of CP basis a_{jk}, one from each "good" projection selected by clustering or some other methods. When the basis vectors $a_{jk}, j \leqslant r$, are nearly orthogonal to each other, the leading singular vector of the selected projected matrix is expected to be reasonably close to one of the CP components, approximating a_{jk} for the same j in all mode k. As the possible directions of randomized projection increase rapidly with dimension d_k, the incoherence condition must decrease with d_k to allow a moderate restart number (i.e. required number of randomized projections) to capture a single CP component. Therefore, the existing incoherence condition in individual tensor modes is hard to avoid in such approaches. Our approach is fundamentally different. As discussed in Section II, the CPCA is designed to take advantage of the multiplicative nature of the higher order coherence.

4) Tucker Models: There exists a large body of work that handles low-rank tensor Tucker decomposition, including [7], [39], [57], [58], [59], [60]. For example, [39] studied HOOI and provides rate optimal statistical bound under Gaussian noise tensor. In the rank-1 case where the CP and Tucker representations are identical, our performance bound in Theorem 6 is equivalent to theirs. Our results and theirs are also in agreement for the estimation of the projection to the column space of CP basis $A_k = (a_{jk}, j \leqslant r)$. The theoretical tool for the analyses of HOOI and our ICO share a similar spirit as both involve projections in the iteration. However, there are several major differences between the statistical analyses in the Tucker and CP models. Moreover, the projection in ICO is very different from previous proposals as discussed in Subsection II-C, thus requiring much more sophisticated analysis. In addition, we develop sharp and useful tensor perturbation bounds in our analysis.

TABLE I

COMPARISON WITH PREVIOUS THEORIES FOR EXISTING CP DECOMPOSITION METHODS WHEN $d_1 \asymp \ldots \asymp d_N \asymp d^{1/N}, \lambda_1 \asymp \lambda_r$ (NEGLECTING LOGARITHMIC FACTORS). HERE CD AND GD ARE COORDINATE DESCENT AND GRADIENT DESCENT, RESPECTIVELY

Algorithms	Incoherence	Iteration complexity	Error (Gaussian Noise)
robust tensor power method [13]	0	$\log(r) + \log\log(1/\epsilon)$	$\sigma\sqrt{d_1}/\lambda_r$
Two-mode HOSVD [42]	0	n/a	$\sigma\sqrt{d_1}/\lambda_r$
randomized projection + power update + CD [29]	$\vartheta_{\max} \lesssim 1/\sqrt{d_1}$	$\log(1/\epsilon)$	$\sigma\sqrt{d_1}/\lambda_r$
spectral method + (vanilla) GD [54]	$\vartheta_{\max} \lesssim 1/\sqrt{d_1}$	$\log(1/\epsilon)$	$\sigma\sqrt{d_1}/\lambda_r$
CPCA + ICO (this paper)	$\delta \wedge (r^{1/2}\vartheta_{\max}^N) \lesssim 1/r^{3/2}$	$\log\log(1/\epsilon)$	$\sigma\sqrt{d_1}/\lambda_r$

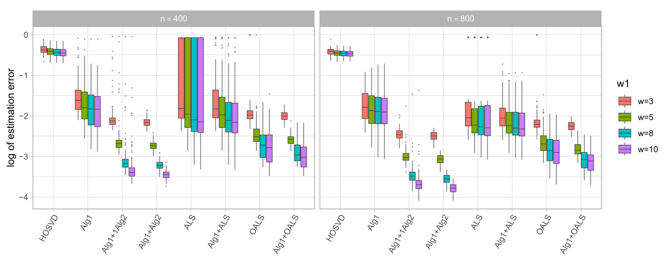

Fig. 1. Boxplots of the logarithm of the estimation error over 100 replications under the spiked covariance tensor setting with $K = 2$ and $\lambda_1 = w_1^2$. The two panels correspond to sample sizes $n = 400, 800$ respectively. The proposed algorithms are labeled as Alg1 (CPCA) and Alg2 (ICO).

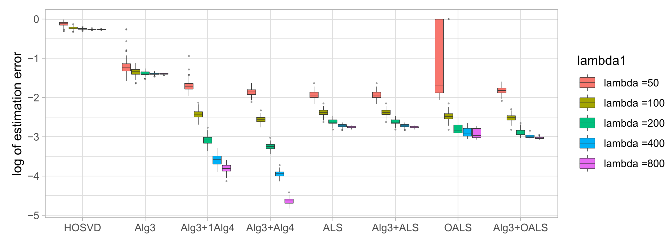

Fig. 2. Boxplots of the logarithm of the estimation error over 100 replications under the low-rank tensor de-noising setting with $N = 4$. The proposed algorithms are labeled as Alg3 (CPCA) and Alg4 (ICO).

IV. NUMERICAL EXPERIMENTS

In this section, we provide some synthetic experiments to compare the performance of the proposed methods, CPCA initialization followed by ICO iterations as in Algorithms 1-4 (Alg1+Alg2 for covariance tensor, Alg3+Alg4 otherwise), with the modified rank one alternating least squares (ALS) [29], orthogonalized alternating least squares (OALS) [48], and higher order SVD (HOSVD). In our simulations, both ALS and OALS use the initialization method proposed in [29] and used in [47] and [15], which applies power and clustering methods to random basis vectors and uses the resulting centroids as initialization. HOSVD, widely used in CP decomposition and tensor completion [54], [56], [61], can be viewed as a baseline initialization method. To better understand CPCA, we also present the results of the method (Alg1 or Alg3) without further improvements and its performance as the initialization of ALS and OALS updates (Alg1-ALS, Alg3-ALS, Alg1-OALS, Alg3-OALS). The estimation error is given by $\max_{j,k} \|\widehat{a}_{jk}\widehat{a}_{jk}^\top - a_{jk}a_{jk}^\top\|_{\rm S}$. The CP basis vectors a_{jk} are first generated independently and uniformly at random from the d_k dimensional unit spherical shell, and then linearly adjusted to satisfy $\max_{i\neq j} |a_{ik}^\top a_{jk}| = 10^{-1/2}$ for order 4 tensors in models (2) and (4).

We first study the finite sample performance with spiked covariance tensors (1). We set $w_{\max}/w_{\min} = 1.25, d_1 = d_2 = 20, r = 3, n = 400, 800$, $K = 2$, $w_{\max} = 3, 5, 8, 10$, so that the covariance tensor is of the order $4 = 2K$. Figure 1 depicts the boxplots of the logarithm of the estimation errors over 100 replicates. In the plot, Alg1+1Alg2 is the one-step ICO estimator after the CPCA initialization. Overall, our method Alg1+Alg2 outperforms all the other methods in all cases. The ICO (Alg2) converges in very few steps, although the number of steps is not reported here. Besides, the one step estimator Alg1+1Alg2 significantly improves over the CPCA initialization (Alg1), and is very close to the final estimator Alg1+Alg2. HOSVD performs much worse than the CPCA initialization (Alg1), probably due to the benefit of multiplicative higher order coherence of the CPCA. The comparisons of ALS against the hybrid Alg1+ALS and OALS against the hybrid Ag1+OALS demonstrate the CPCA as a better method than clustering or other randomized screening methods for initialization, although the CPCA initialization (Alg1) standing alone may perform worse than iterative methods (slightly so compared with ALS and more clearly so with OALS). In fact the hybrid methods with the CPCA initialization improve the original randomized initialized ALS and OALS significantly, especially when the signal strength $w_{\max}$ is large.

We also explore our methods under the low-rank tensor de-noising setting (4). We consider a 4-way tensor with $d_1 = d_2 = d_3 = d_4 = 20$, $\lambda_{\max}/\lambda_{\min} = 1.25, r = 3$, and $\lambda_{\max} = 50, 100, 200, 400, 800$. Figure 2 quantifies the performance of different algorithms in terms of the logarithm of the estimation errors. Except for $\lambda_{\max} = 50$, Alg3+Alg4 is superior to all the other algorithms. When $\lambda_{\max} = 50$, ALS and Alg3+ALS are slightly better than Alg3+Alg4 and Alg3+OALS. Again, HOSVD underperforms the CPCA initialization (Alg3). Figure 2 also shows the benefits of one step estimator Alg3+1Alg4. Although Alg3+ALS has similar behavior as ALS in this setting, we do not need to generate a large number of random initialization in the hybrid method Alg3+ALS.

Next, we consider two additional cases of order 6 tensors in models (2) and (4) with basis vectors satisfying

TABLE II

RUN TIME FOR DIFFERENT INITIALIZATION METHODS OVER 100 REPLICATIONS UNDER THE SPIKED COVARIANCE TENSOR SETTING WITH $K = 2$. HERE RUN TIME IS THE MEAN AND STANDARD DEVIATIONS OF THE RUN TIME IN SECONDS. ALS-INIT USES 30 RESTART NUMBERS

Algorithms	$d_1 = d_2 = 20$	$d_1 = d_2 = 30$	$d_1 = d_2 = 40$	$d_1 = d_2 = 50$	$d_1 = d_2 = 60$
HOSVD	$0.12_{(0.02)}$	$1.18_{(0.28)}$	$4.62_{(0.89)}$	$12.58_{(2.27)}$	$27.98_{(5.37)}$
ALS-init	$6.97_{(1.93)}$	$38.95_{(8.21)}$	$133.20_{(25.78)}$	$332.68_{(81.74)}$	$726.75_{(177.53)}$
Alg1 (CPCA)	$0.19_{(0.02)}$	$2.07_{(0.28)}$	$10.10_{(1.13)}$	$33.71_{(2.96)}$	$94.55_{(9.41)}$

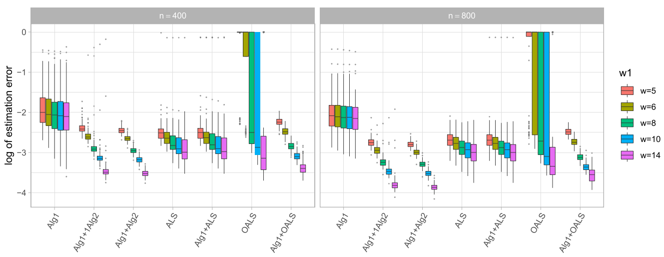

Fig. 3. Boxplots of the logarithm of the estimation error over 100 replications under the spiked covariance tensor setting with $K = 3$ and $\lambda_1 = w_1^2$. Two panels correspond to two sample sizes $n = 400, 800$. The proposed algorithms are labeled as Alg1 (CPCA) and Alg2 (ICO).

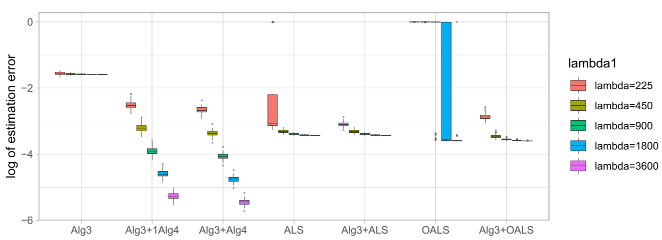

Fig. 4. Boxplots of the logarithm of the estimation error over 100 replications under the low-rank tensor denoising setting with $N = 6$. The proposed algorithms are labeled as Alg3 (CPCA) and Alg4 (ICO).

$\max_{i \neq j} |a_{ik}^\top a_{jk}|^3 = 0.1$. In a spiked covariance tensor setting (1), we set $w_{\max}/w_{\min} = 1.25, r = 3, d_1 = d_2 = d_3 = 20, n = 400, 800, K = 3$, and $w_{\max} = 5, 6, 8, 10, 14$. In the low-rank tensor denoising setting (4), we set $d_k = 20, 1 \leqslant k \leqslant 6, \lambda_{\max}/\lambda_{\min} = 1.25, r = 3$, and $\lambda_{\max} = 225, 450, 900, 1800, 3600$. We omit HOSVD as it is always much worse than the CPCA initialization. The results are similar to order 4 tensors. From Figure 3, Alg1+Alg2 are the best one in all cases. The advantages are more obvious when $w_{\max}$ is large. OALS with randomized initialization has a great deal of variabilities, which can be significantly improved by the CPCA initialization (Alg1+OALS). Though ALS and Alg1+ALS have almost the same performance, Alg1+ALS does not require a large number of random initialization. The results in the tensor denoise setting, reported in Figure 4, are similar to those in the spiked covariance tensor model setting in Figure 3, except the case $\lambda_{\max} = 225, 450$. Alg3+ALS fares better than the other approaches for $\lambda_{\max} = 225$, while Alg3+OALS is the best for $\lambda_{\max} = 450$. Although the proposed algorithms do not always outperform ALS and OALS, they underperform only slightly and in very few simulation configurations and they are faster and easier to implement. Moreover, the simulation results demonstrate that the CPCA initialization is superior to the randomized initialization with ALS and OALS.

To evaluate the computational cost of different initialization methods, we also report the run time of the CPCA initialization, HOSVD, and the randomized initialization in [29] (ALS-init) under a spiked covariance tensor setting (1). We set $w_{\max}/w_{\min} = 1.25, r = 3, n = 800, K = 2, w_{\max} = 10$, and vary $d_1 = d_2 = 20, 30, 40, 50, 60$. From Table II, it can be seen that ALS-init requires much longer run time for each simulation than the other methods. The reason may be that ALS-init needs a large number of restarts to recover all the CP basis. Meanwhile, HOSVD has significantly shorter run time than the CPCA, and the ratio of the costs seems stable as the dimension increases. Thus, the far superior performance of the CPCA justifies its (still manageable) computational costs compared with HOSVD.

In summary, the proposed Algorithms 1-4 are more accurate than existing methods in the simulation experiments in general. Algorithms 1 and 3 can also be a superior initialization to plug in existing algorithms, and is faster and much simpler to implement than randomized initializations. It is worth noting that in the case of order 6 tensors where the incoherence $\max_{i \neq j} |a_{ik}^\top a_{jk}| = 10^{-1/3}$ is larger, both ALS and OALS perform poorly, while the proposed methods still work well.

V. FINAL REMARK

In this paper, we propose new initialization (CPCA) and refinement (ICO) algorithms for tensor CP decomposition of high dimensional non-orthogonal spike tensors. Our methods tolerate a higher level of coherence among the basis vectors (a_{jk}), and achieve faster computational convergence rate and sharper statistical error bounds, compared with existing methods. The proposed methods are applicable to a broad class of structured tensors, including the spiked covariance tensors (2) and general noisy high order tensors (4). In particular, our proposed algorithms show stable convergence and exhibit pronounced advantage especially as the order of the tensor increases. Numerical studies display empirically favorable performance of the proposed methods.

APPENDIX A
ANALYSIS IN THE NOISELESS CASE: MATRIX AND TENSOR PERTURBATION BOUNDS

This section provides the analysis of the CPCA and ICO algorithms in the noiseless case with $\Psi = 0$ in models (2) and (4). The results in this section are dimension free in the sense that their conditions and conclusions depend only on the angles among the basis vectors and their estimates and the principle angles among spaces, not on d_k. We first present the proofs of Propositions 1, 2, 3 and 4. These propositions provide a road map of the proof of Theorem 1, which is to follow, and some general techniques to study model (2). Then, we present the proofs of Propositions 5 and 6. Propositions 1, 5, 3 and 6 provide a road map of the proof of Theorem 4 at the end of this section and some general techniques to study model (4). For readers' convenience, we restate the propositions and theorems before their proofs.

Proposition 1: For any set S of tensor modes, define $a_{jS} = \mathrm{vec}(\otimes_{k\in S} a_{jk})$, $A_S = (a_{1S}, \ldots, a_{rS})$, $\vartheta_S = \max_{1\leqslant i<j\leqslant r} |a_{iS}^\top a_{jS}|$ and $\delta_S = \|A_S^\top A_S - I_r\|_{\mathrm{S}}$. Define

$$\mu_S = \max_j \min_{k_1,k_2\in S} \max_{i\neq j} \prod_{k\neq k_1, k\neq k_2, k\in S} \sqrt{r}|\sigma_{ij,k}|/\eta_{jk}.$$

as the (leave-two-out) mutual coherence of $\{A_j, j\in S\}$. Then, $\mu_S \in [1, r^{|S|/2-1}]$,

$$\delta_S \leqslant \min_{k\in S} \delta_k, \quad \delta_S \leqslant (r-1)\vartheta_S \leqslant (r-1)\prod_{k\in S}\vartheta_k, \tag{9}$$

$$\delta_S \leqslant \mu_S r^{1-|S|/2} \max_{j\leqslant r} \prod_{k\in S} \eta_{jk} \leqslant \mu_S r^{1-|S|/2} \prod_{k\in S} \delta_k. \tag{10}$$

When $S = [K]$, the above inequalities hold with $\{\delta_S, \vartheta_S\}$ replaced by the $\{\delta, \vartheta\}$ in (8).

Proof of Proposition 1: For notational simplicity, we only prove the case $S = [K]$, as the extension to general S is straightforward. Recall that $\delta = \|A^\top A - I_r\|_{\mathrm{S}}$ and $\delta_k = \|A_k^\top A_k - I_r\|_{\mathrm{S}}$. Because $A^\top A = (A_1^\top A_1)\circ\cdots\circ(A_K^\top A_K)$ is the Hadamard product of correlation matrices, the spectrum of $A^\top A$ is contained inside the spectrum limits of $A_k^\top A_k$ for each k, so that

$$\delta \leqslant \min_{1\leqslant k\leqslant K} \delta_k.$$

Because $A^\top A - I_r$ is symmetric, its spectrum norm is bounded by its ℓ_1 norm,

$$\delta \leqslant \max_{j\leqslant r} \textstyle\sum_{i\neq j} |a_i^\top a_j| \leqslant (r-1)\vartheta \leqslant (r-1)\displaystyle\prod_{k=1}^K \vartheta_k$$

due to $|a_i^\top a_j| = \prod_{k=1}^K |a_{ik}^\top a_{jk}| = \prod_{k=1}^K |\sigma_{ij,k}|$. Moreover, for any $j\leqslant r$ and $1\leqslant k_1 < k_2 \leqslant K$,

$$\begin{aligned}&\sum_{i\neq j}\prod_{k=1}^K |\sigma_{ij,k}| \\ &\leqslant \sum_{i\neq j} |\sigma_{ij,k_1}\sigma_{ij,k_2}| \max_{i\neq j} \prod_{k\neq k_1, k\neq k_2} |\sigma_{ij,k}| \\ &\leqslant \Big(\textstyle\prod_{k=1}^K \eta_{jk}\Big) r^{-(K-2)/2} \displaystyle\max_{i\neq j} \prod_{k\neq k_1, k\neq k_2} \sqrt{r}|\sigma_{ij,k}|/\eta_{jk},\end{aligned}$$

as $\eta_{jk} = (\sum_{i\neq j}\sigma_{ij,k}^2)^{1/2}$. The proof is complete as k_1 and k_2 are arbitrary. □

Proposition 2: Let $d \geqslant r$ and $A \in \mathbb{R}^{d\times r}$ with $\|A^\top A - I_r\|_{\mathrm{S}} \leqslant \delta$. Let $A = \widetilde{U}_1\widetilde{D}_1\widetilde{U}_2^\top$ be the SVD of A, and $U = \widetilde{U}_1\widetilde{U}_2^\top$. Then, $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \leqslant \delta\|\Lambda\|_{\mathrm{S}}$ for all nonnegative-definite matrices Λ in $\mathbb{R}^{r\times r}$.

Proof of Proposition 2: An extension of Proposition 2, Proposition 5, is proved later. □

Proposition 3: Let $M \in \mathbb{R}^{d_1\times d_2}$ be a matrix with $\|M\|_{\mathrm{F}} = 1$ and a and b be unit vectors respectively in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$. Let $\widehat{a}$ be the top left singular vector of M. Then,

$$\begin{aligned}&\big(\|\widehat{a}\widehat{a}^\top - aa^\top\|_{\mathrm{S}}^2\big)\wedge(1/2) \\ &\leqslant \|\mathrm{vec}(M)\mathrm{vec}(M)^\top - \mathrm{vec}(ab^\top)\mathrm{vec}(ab^\top)^\top\|_{\mathrm{S}}^2.\end{aligned} \tag{14}$$

Proof of Proposition 3. Let $\sum_{j=1}^r \sigma_j u_j v_j^\top$ be the SVD of M with singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_r$ where r is the rank of M. Because $\mathrm{vec}(u_j v_j^\top)$ are orthonormal in $\mathbb{R}^{d_1 d_2}$,

$$\mathrm{vec}(M)^\top \mathrm{vec}(ab^\top) = a^\top M b = \sum_{j=1}^r \sigma_j (u_j^\top a)(v_j^\top b)$$

with $\sum_{j=1}^r \sigma_j^2 = \|M\|_{\mathrm{F}}^2 = 1$, $\sum_{j=1}^r (u_j^\top a)^2 \leqslant \|a\|_2^2 = 1$ and $\sum_{j=1}^r (v_j^\top b)^2 \leqslant \|b\|_2^2 = 1$. Because $\sigma_1 \geqslant \cdots \geqslant \sigma_r$,

$$\big|a^\top M b\big| \leqslant \sigma_1 \Big(\sum_{j=1}^r (u_j^\top a)^2\Big)^{1/2}\Big(\sum_{j=1}^r (v_j^\top b)^2\Big)^{1/2} = \sigma_1$$

Similarly, by Cauchy-Schwarz,

$$\begin{aligned}\big|a^\top M b\big|^2 &\leqslant \sum_{j=1}^r \sigma_j^2 (u_j^\top a)^2 \\ &\leqslant \sigma_1^2 (u_1^\top a)^2 + \big(1-\sigma_1^2\big)\big(1-(u_1^\top a)^2\big).\end{aligned} \tag{37}$$

When $(u_1^\top a)^2 \geqslant 1/2$, the maximum on the right-hand side above is achieved at $\sigma_1^2 = 1$, so that $\big|a^\top M b\big|^2 \leqslant (u_1^\top a)^2$; Otherwise, the right-hand side of (37) is maximized at $\sigma_1^2 = \big|a^\top M b\big|^2$, so that $\big|a^\top M b\big|^2 \leqslant 1 - \big|a^\top M b\big|^2$. Thus, $\big|a^\top M b\big|^2 > 1/2$ implies $\big|a^\top M b\big|^2 \leqslant (u_1^\top a)^2$. By (5), this is equivalent to (14). □

Proposition 4: Let $T^* = \mathbb{E}[T]$ with the tensor T in (2). Given $\widetilde{A}_\ell = (\widetilde{a}_{1\ell}, \ldots, \widetilde{a}_{r\ell}), \ell \in [K]\backslash\{k\}$, let $\widetilde{a}_{jk}^*$ be the top eigenvector of $\widetilde{T}_{jk}^* = T^* \times_{l\in[2K]\backslash\{k,K+k\}} \widetilde{b}_{jl}^\top \in \mathbb{R}^{d_k\times d_k}$ with the $\widetilde{b}_{jl}$ in (16). Then,

$$\begin{aligned}&\|a_{jk}a_{jk}^\top - \widetilde{a}_{jk}^*\widetilde{a}_{jk}^{*\top}\|_{\mathrm{S}} \\ &\leqslant 2\big(1+\delta_k\big)\big(\lambda_1/\lambda_j\big)\prod_{\ell\in[K]\backslash\{k\}} (\widetilde{\phi}_\ell/(1-\widetilde{\phi}_\ell)_+)^2,\end{aligned}$$

where $\widetilde{\phi}_\ell = \widetilde{\psi}_\ell/\big(\sqrt{(1-\delta_\ell)(1-1/(4r))} - \sqrt{r}\widetilde{\psi}_\ell\big)_+$ with $\widetilde{\psi}_\ell = \max_{j\leqslant r}\big\|\widetilde{a}_{j\ell}\widetilde{a}_{j\ell}^\top - a_{j\ell}a_{j\ell}^\top\big\|_{\mathrm{S}}$.

Proof of Proposition 4: For any diagonal matrix Π_r with $\Pi_r^2 = I_r$, $\widetilde{A}_\ell\Pi_r((\widetilde{A}_\ell\Pi_r)^\top\widetilde{A}_\ell\Pi_r)^{-1} = \widetilde{A}_\ell(\widetilde{A}_\ell^\top\widetilde{A}_\ell)^{-1}\Pi_r$. Thus, because $\widetilde{T}_{jk}^*$ does not depend on the signs of $\widetilde{b}_{j\ell} = \widetilde{b}_{j,\ell+K}$, we assume without loss of generality that $\widetilde{a}_{i\ell}^\top a_{i\ell} \geqslant 0$ for all i and ℓ. Let $\widetilde{\Sigma}_\ell = \widetilde{A}_\ell^\top\widetilde{A}_\ell$. Assume without loss of generality that $r\widetilde{\psi}_\ell^2/(1-1/(4r)) \leqslant 1$, so that $2(1-(1-\widetilde{\psi}_\ell^2)^{1/2}) \leqslant \widetilde{\psi}_\ell^2/(1-1/(4r))$. Consequently,

$$\begin{aligned}\|\widetilde{A}_\ell - A_\ell\|_{\mathrm{S}}^2 &\leqslant r\max_{j\leqslant r}\big\|\widetilde{a}_{j\ell} - a_{j\ell}\big\|_2^2 = 2r(1-(1-\widetilde{\psi}_\ell^2)^{1/2}) \\ &\leqslant r\widetilde{\psi}_\ell^2/(1-1/(4r))\end{aligned}$$

As $\widetilde{b}_{j\ell} = \widetilde{A}_\ell(\widetilde{A}_\ell^\top \widetilde{A}_\ell)^{-1}e_j$, $\|\widetilde{b}_{j\ell}\|_2 = \{e_j^\top(\widetilde{A}_\ell^\top \widetilde{A}_\ell)^{-1}e_j\}^{1/2} \leqslant \max_{\|u\|_2=1}\|\widetilde{A}_\ell u\|_2^{-1}$, so that

$$\begin{aligned}&\max_{i\leqslant r}\|\widetilde{a}_{i\ell}-a_{i\ell}\|_2\|\widetilde{b}_{j\ell}\|_2\\ &\leqslant\big(\widetilde{\psi}_\ell/\sqrt{1-1/(4r)}\big)/\big(\sqrt{1-\delta_\ell}-r^{1/2}\widetilde{\psi}_\ell/\sqrt{1-1/(4r)}\big)_+\\ &=\widetilde{\phi}_\ell.\end{aligned}$$

Let $w_{ij,\ell} = a_{i\ell}^\top \widetilde{b}_{j\ell}/a_{j\ell}^\top \widetilde{b}_{j\ell}$, $v_{i,jk} = (\lambda_i/\lambda_j)\prod_{\ell\in[K]\setminus\{k\}} w_{ij,\ell}^2$, $v_{jk} \in \mathbb{R}^r$ be the vector with elements $v_{i,jk}$, and $\widetilde{\lambda}_j = \lambda_j\prod_{\ell\in[K]\setminus\{k\}}(a_{j\ell}^\top\widetilde{b}_{j\ell})^2$. As $\widetilde{a}_{i\ell}^\top\widetilde{b}_{j\ell} = I_{\{i=j\}}$, for $i\neq j$,

$$\begin{aligned}|w_{ij,\ell}| &= \frac{|(a_{i\ell}-\widetilde{a}_{i\ell})^\top\widetilde{b}_{j\ell}|}{|1+(a_{j\ell}-\widetilde{a}_{j\ell})^\top\widetilde{b}_{j\ell}|} \leqslant \frac{\widetilde{\phi}_\ell}{1-\widetilde{\phi}_\ell},\\ |v_{i,jk}| &\leqslant (\lambda_1/\lambda_j)\bigg(\prod_{\ell\in[K]\setminus\{k\}}\frac{\widetilde{\phi}_\ell}{1-\widetilde{\phi}_\ell}\bigg)^2.\end{aligned}$$

As $\widetilde{T}_{jk}^*/\widetilde{\lambda}_j = \sum_{i=1}^r a_{ik}a_{ik}^\top v_{i,jk}$ and eigenvectors do not depend in scaling,

$$\begin{aligned}\|a_{jk}a_{jk}^\top - \widetilde{a}_{jk}^*\widetilde{a}_{jk}^{*\top}\|_{\mathrm{S}} &\leqslant 2\|\textstyle\sum_{i\neq j}^r a_{ik}a_{ik}v_{i,jk}\|_{\mathrm{S}}\\ &\leqslant 2\|A_k\|_{\mathrm{S}}^2\max_{i\neq j}v_{i,jk}\end{aligned}$$

by Wedin's theorem [62]. The conclusion follows. $\square$

Theorem 1: Suppose Algorithm 1 (CPCA) is applied to the noiseless $T^* = \sum_{j=1}^r \lambda_j \otimes_{k=1}^{2K} a_{jk}$ with $a_{j,K+k} = a_{jk}$. Then, (15) holds for the resulting $\widehat{\lambda}_j^{\mathrm{cpca}}$ and $\widehat{a}_{jk}^{\mathrm{cpca}}$. Let $\lambda_{\min,\pm} = \min_{1\leqslant j\leqslant r}\lambda_{j,\pm}$ be the minimum eigengap. Suppose further that

$$\begin{aligned}2\max\big\{\delta_{\max},(\sqrt{r}+1)\psi_0\big\}&\leqslant 1,\\ 3(\lambda_1/\lambda_r)\psi_0^{2K-3}&\leqslant\rho<1,\end{aligned}\tag{17}$$

where $\delta_{\max} = \max_{k\leqslant K}\delta_k$ with the δ_k in (6) and $\psi_0 = (1+2\lambda_1/\lambda_{\min,\pm})\delta$ with the δ in (8). Let $\gamma_K \in (3-3/K, 3)$ be the solution of $\gamma_K^K - 3\gamma_K^{K-1}+2 = 0$, e.g. $\gamma_3 = 2.732$, $\gamma_4 = 2.919$. If the resulting $\widehat{a}_{jk}^{\mathrm{cpca}}$ are used as the initialization of Algorithm 2 (ICO) with the same data T^*, then

$$\max_{j\leqslant r}\big\|\widehat{a}_{jk}^{(m)}\widehat{a}_{jk}^{(m)\top} - a_{jk}a_{jk}^\top\big\|_{\mathrm{S}} \leqslant \psi_{m,k} = \psi_0\rho^{\gamma_K^{(m-1)K+k-1}}$$

and $\max_{1\leqslant k\leqslant K}\psi_{m,k} \leqslant \epsilon$ within m iterations, where $m = \lceil K^{-1}\{1+(\log\gamma_K)^{-1}\log(\log(\psi_0/\epsilon)/\log(1/\rho))\}\rceil$.

Proof of Theorem 1: Let $U = (u_1,\ldots,u_r)$ be the orthonormal matrix corresponding to $A = (a_1,\ldots,a_r)$ as in Proposition 2 where $a_j = \mathrm{vec}(\otimes_{k=1}^K a_{jk})$. Let $\mathrm{mat}_{[K]}(T^*) = \sum_{j=1}^r \widehat{\lambda}_j^{\mathrm{cpca}}\widehat{u}_j\widehat{u}_j^\top$ be the eigenvalue decomposition as in (11). By Proposition 2 and Wedin's perturbation theorem,

$$\begin{aligned}\|\widehat{u}_j\widehat{u}_j^\top - a_ja_j^\top\|_{\mathrm{S}} &\leqslant \|u_ju_j^\top - a_ja_j^\top\|_{\mathrm{S}} + \|\widehat{u}_j\widehat{u}_j^\top - u_ju_j^\top\|_{\mathrm{S}}\\ &\leqslant \delta + 2(\lambda_1\delta)/\lambda_{j,\pm} \leqslant \psi_0\end{aligned}$$

and $|\widehat{\lambda}_j^{\mathrm{cpca}} - \lambda_j| \leqslant \delta\lambda_1$. Thus, (15) follows from Proposition 3. Moreover, under (18) we have $\psi_0 < 1/(\sqrt{r}+1) \leqslant 1/2$, so that (15) yields $\max_{j\leqslant r}\big\|\widehat{a}_{j\ell}^{\mathrm{cpca}}\widehat{a}_{j\ell}^{\mathrm{cpca}\top} - a_{j\ell}a_{j\ell}^\top\big\|_{\mathrm{S}} \leqslant \psi_0$. Now define $\psi'_{m,k} = \max_{j\leqslant r}\big\|\widehat{a}_{jk}^{(m)}\widehat{a}_{jk}^{(m)\top} - a_{jk}a_{jk}^\top\big\|_{\mathrm{S}}$ with $\widehat{a}_{jk}^{(0)} = \widehat{a}_{jk}^{\mathrm{cpca}}$. By Proposition 4 and (18), $\psi'_{1,1} \leqslant \rho\psi_0$ and this would contribute the extra factor ρ twice in the application of Proposition 4 to $\psi'_{1,2}$, resulting in $\psi'_{1,2} \leqslant \rho^3\psi_0$, so on and so forth. In general, $\psi'_{m,k} \leqslant \rho^{n_{(m-1)K+k}}\psi_0$ with $n_1 = 1$, $n_2 = 3,\ldots,n_K = 3^{K-1}$, and $n_{k+1} = 1+2\sum_{\ell=1}^{K-1}n_{k+1-\ell}$ for $k > K$. As $2(1-\gamma_K^{-K+1}) = \gamma_K - 1$, by induction for $k \geqslant K$,

$$n_{k+1} \geqslant 2\big(\gamma_K^{k-1}+\cdots+\gamma_K^{k-K+1}\big) = \gamma_K^k\frac{2(1-\gamma_K^{-K+1})}{\gamma_K-1} = \gamma_K^k.$$

The function $f(\gamma) = \gamma^K - 3\gamma^{K-1}+2$ is decreasing in $(1, 3-3/K)$ and increasing $(3-3/K,\infty)$. Because $f(1) = 0$ and $f(3) = 2 > 0$, we have $3-3/K < \gamma_K < 3$. $\square$

Proposition 5: Let $A \in \mathbb{R}^{d_1\times r}$ and $B \in \mathbb{R}^{d_2\times r}$ with $\|A^\top A - I_r\|_{\mathrm{S}} \vee \|B^\top B - I_r\|_{\mathrm{S}} \leqslant \delta$ and $d_1 \wedge d_2 \geqslant r$. Let $A = \widetilde{U}_1\widetilde{D}_1\widetilde{U}_2^\top$ be the SVD of A, $U = \widetilde{U}_1\widetilde{U}_2^\top$, $B = \widetilde{V}_1\widetilde{D}_2\widetilde{V}_2^\top$ the SVD of B, and $V = \widetilde{V}_1\widetilde{V}_2^\top$. Then, $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \leqslant \delta\|\Lambda\|_{\mathrm{S}}$ for all nonnegative-definite matrices Λ in $\mathbb{R}^{r\times r}$, and $\|AQB^\top - UQV^\top\|_{\mathrm{S}} \leqslant \sqrt{2}\delta\|Q\|_{\mathrm{S}}$ for all $r\times r$ matrices Q.

Proof of Proposition 5: Let $A = \widetilde{U}_1\widetilde{D}_1\widetilde{U}_2^\top$ and $B = \widetilde{V}_1\widetilde{D}_2\widetilde{V}_2^\top$ be respectively the SVD of A and B with $\widetilde{D}_1 = \mathrm{diag}(\widetilde{\sigma}_{11},\ldots,\widetilde{\sigma}_{1r})$ and $\widetilde{D}_2 = \mathrm{diag}(\widetilde{\sigma}_{21},\ldots,\widetilde{\sigma}_{2r})$. Let $U = \widetilde{U}_1\widetilde{U}_2^\top$ and $V = \widetilde{V}_1\widetilde{V}_2^\top$. We have $\|\widetilde{D}_1^2 - I_r\|_{\mathrm{S}} = \|A^\top A - I_r\|_{\mathrm{S}} \leqslant \delta$ and $\|\widetilde{D}_2^2 - I_r\|_{\mathrm{S}} = \|B^\top B - I_r\|_{\mathrm{S}} \leqslant \delta$. Moreover,

$$\begin{aligned}&\|AQB^\top - UQV^\top\|_{\mathrm{S}}^2\\ &=\max\nolimits_{\|u_1\|_2=\|u_2\|_2=1}\big|u_1^\top\big(\widetilde{D}_1\widetilde{U}_2^\top Q\widetilde{V}_2\widetilde{D}_2 - \widetilde{U}_2^\top Q\widetilde{V}_2\big)u_2\big|^2\\ &\leqslant 2\|Q\|_{\mathrm{S}}^2\max\nolimits_{\|u_1\|_2=\|u_2\|_2=1}\big\|\widetilde{D}_2u_2u_1^\top\widetilde{D}_1 - u_2u_1^\top\big\|_{\mathrm{F}}^2\\ &=2\|Q\|_{\mathrm{S}}^2\max_{\|u_1\|_2=\|u_2\|_2=1}\sum_{i=1}^r\sum_{j=1}^r u_{1i}^2u_{2j}^2\big(\widetilde{\sigma}_{1i}\widetilde{\sigma}_{2j}-1\big)^2\end{aligned}$$

with $u_\ell = (u_{\ell 1},\ldots,u_{\ell r})^\top$, $\sqrt{(1-\delta)_+} \leqslant \widetilde{\sigma}_{\ell j} \leqslant \sqrt{1+\delta}$, $\ell = 1,2$. The maximum on the right-hand side above is attained at $\widetilde{\sigma}_{\ell j} = \sqrt{(1-\delta)_+}$ or $\sqrt{1+\delta}$ by convexity. As $(\sqrt{(1-\delta)_+}\sqrt{1+\delta}-1)^2 \leqslant \delta^4\wedge 1$, we have $\|AQB^\top - UQV^\top\|_{\mathrm{S}}^2 \leqslant 2\|Q\|_{\mathrm{S}}^2\delta^2$. For nonnegative-definite Λ and $B = A$, $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} = \|\widetilde{D}_1\widetilde{U}_2^\top\Lambda\widetilde{U}_2\widetilde{D}_1 - \widetilde{U}_2^\top\Lambda\widetilde{U}_2\|_{\mathrm{S}}$ and

$$\begin{aligned}&\big|u^\top\big(\widetilde{D}_1\widetilde{U}_2^\top\Lambda\widetilde{U}_2\widetilde{D}_1 - \widetilde{U}_2^\top\Lambda\widetilde{U}_2\big)u\big| = \bigg|\sum_{j=1}^2\tau_jv_j^\top\widetilde{U}_2^\top\Lambda\widetilde{U}_2v_j\bigg|\\ &\leqslant\begin{cases}\|\Lambda\|_{\mathrm{S}}(|\tau_1|\vee|\tau_2|), & \tau_1\tau_2<0,\\ \|\Lambda\|_{\mathrm{S}}(|\tau_1+\tau_2|), & \tau_1\tau_2\geqslant 0,\end{cases}\end{aligned}$$

where $\sum_{j=1}^2\tau_jv_jv_j^\top$ is the eigenvalue decomposition of $\widetilde{D}_1uu^\top\widetilde{D}_1 - uu^\top$. Similar to the general case, $(|\tau_1|\vee|\tau_2|)^2 \leqslant \tau_1^2+\tau_2^2 = \big\|\widetilde{D}_1uu^\top\widetilde{D}_1 - uu^\top\big\|_{\mathrm{F}}^2 \leqslant \delta^2$ and $|\tau_1+\tau_2| = |\mathrm{tr}(\widetilde{D}_1uu^\top\widetilde{D}_1 - uu^\top)| \leqslant \|\widetilde{D}_1\widetilde{D}_1 - I_r\|_{\mathrm{S}} \leqslant \delta$. Hence, $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \leqslant \|\Lambda\|_{\mathrm{S}}\delta$. $\square$

Proposition 6: Let $T^* = \mathbb{E}[T]$ with the tensor T in (4). Given $\widetilde{A}_\ell = (\widetilde{a}_{1\ell},\ldots,\widetilde{a}_{r\ell})$, $\ell\in[N]\setminus\{k\}$, let $(\widetilde{b}_{1\ell},\ldots,\widetilde{b}_{r\ell}) = \widetilde{A}_\ell(\widetilde{A}_\ell^\top\widetilde{A}_\ell)^{-1}$, $\widetilde{T}_{jk}^* = T^*\times_{l\in[N]\setminus\{k\}}\widetilde{b}_{jl}^\top \in \mathbb{R}^{d_k}$, $\widetilde{a}_{jk}^* = \widetilde{T}_{jk}^*/\|\widetilde{T}_{jk}^*\|_2$ and $\widetilde{\lambda}_j^* = T^*\times_{l\in[N]}\widetilde{b}_{jl}^\top$. Then,

$$\begin{aligned}2-2\big|a_{jk}^\top\widetilde{a}_{jk}^*\big| &\leqslant 2(r-1)\big(1+\delta_k\big)\bigg(\frac{\lambda_1}{\lambda_j}\prod_{\ell\in[N]\setminus\{k\}}\frac{\widetilde{\phi}_\ell}{1-\widetilde{\phi}_\ell}\bigg)^2,\\ \big|\widetilde{\lambda}_j^*/\lambda_j - 1\big| &\leqslant \sum_{\ell=1}^N\widetilde{\phi}_\ell + (r-1)(\lambda_1/\lambda_j)\prod_{\ell=1}^N\widetilde{\phi}_\ell,\end{aligned}$$

where $\widetilde{\phi}_\ell = \widetilde{\psi}_\ell / \big(\sqrt{1-\delta_\ell} - \sqrt{r}\widetilde{\psi}_\ell\big)_+$ with $\widetilde{\psi}_\ell = \max_{j\leqslant r}\big(2 - 2\big|\widetilde{a}_{j\ell}^\top a_{j\ell}\big|\big)^{1/2}$.

Proof of Proposition 6: By the argument in the beginning of the proof of Proposition 4, the conclusion of Proposition 6 does not depend on the signs of $\widetilde{a}_{j\ell}$. Thus, we assume without loss of generality that $\widetilde{a}_{i\ell}^\top a_{i\ell} \geqslant 0$ for all i and ℓ. Instead of $\max_{j\leqslant r}\big\|\widetilde{a}_{j\ell} - a_{j\ell}\big\|_2^2 \leqslant \widetilde{\psi}_\ell^2/(1-1/(4r))$ in the proof of Proposition 4, we have the simpler $\max_{j\leqslant r}\big\|\widetilde{a}_{j\ell} - a_{j\ell}\big\|_2 \leqslant \widetilde{\psi}_\ell$ here. Modifying the proof there accordingly, we have $\max_{i\leqslant r}\big\|\widetilde{a}_{i\ell} - a_{i\ell}\big\|_2 \|\widetilde{b}_{j\ell}\|_2 \leqslant \widetilde{\psi}_\ell / \big(\sqrt{1-\delta_\ell} - r^{1/2}\widetilde{\psi}_\ell\big)_+ = \widetilde{\phi}_\ell$. Again let $w_{ij,\ell} = a_{i\ell}^\top \widetilde{b}_{j\ell} / a_{j\ell}^\top \widetilde{b}_{j\ell}$, $v_{i,jk} = (\lambda_i/\lambda_j)\prod_{\ell\in[N]\backslash\{k\}} w_{ij,\ell}$, $v_{jk} \in \mathbb{R}^r$ be the vector with elements $v_{i,jk}$, and $\widetilde{\lambda}_j = \lambda_j \prod_{\ell\in[N]\backslash\{k\}}(a_{j\ell}^\top \widetilde{b}_{j\ell})$. As $\widetilde{a}_{i\ell}^\top \widetilde{b}_{j\ell} = I_{\{i=j\}}$, for $i \neq j$,

$$|w_{ij,\ell}| = \frac{|(a_{i\ell} - \widetilde{a}_{i\ell})^\top \widetilde{b}_{j\ell}|}{|1 + (a_{j\ell} - \widetilde{a}_{j\ell})^\top \widetilde{b}_{j\ell}|} \leqslant \frac{\widetilde{\phi}_\ell}{1-\widetilde{\phi}_\ell},$$

$$\big|v_{i,jk}\big| \leqslant (\lambda_1/\lambda_j) \prod_{\ell\in[N]\backslash\{k\}} \left(\frac{\widetilde{\phi}_\ell}{1-\widetilde{\phi}_\ell}\right).$$

As $\widetilde{T}_{jk}^*/\widetilde{\lambda}_j = \sum_{i=1}^r a_{ik} v_{i,jk}$ and $v_{j,jk} = 1$,

$$\begin{aligned}&\big\|\widetilde{T}_{jk}^*/\widetilde{\lambda}_j - a_{jk}\big\|_2^2\\ =&\textstyle\sum_{i_1\in[r]\backslash\{j\}}\sum_{i_2\in[r]\backslash\{j\}} \sigma_{i_1 i_2,k} v_{i_1,jk} v_{i_2,jk}\\ \leqslant&(r-1)\big(1+\delta_k\big)(\lambda_1/\lambda_j)^2 \prod_{\ell\in[N]\backslash\{k\}} (\widetilde{\phi}_\ell/(1-\widetilde{\phi}_\ell))^2.\end{aligned}$$

Let 2θ be the angle between a_{jk} and $\widetilde{a}_{jk}^* = \widetilde{T}_{jk}^*/\|\widetilde{T}_{jk}^*\|_2$. We have $2\big(1 - a_{jk}^\top \widetilde{a}_{jk}^*\big) = \|a_{jk} - \widetilde{a}_{jk}^*\|_2^2 = (2\sin\theta)^2 = 2(1-\cos(2\theta)) \leqslant 2(1-\cos^2(2\theta)) = 2\sin^2(2\theta) \leqslant 2\big\|\widetilde{T}_{jk}^*/\widetilde{\lambda}_j - a_{jk}\big\|_2^2$.

Similarly, as $\widetilde{\lambda}_j^* - \lambda_j = \lambda_j(\prod_{\ell\in[N]} a_{j\ell}^\top \widetilde{b}_{j\ell} - 1) + \sum_{i\in[r]\backslash\{j\}}\lambda_i \prod_{\ell\in[N]} a_{i\ell}^\top \widetilde{b}_{j\ell}$, we have

$$\big|\widetilde{\lambda}_j^*/\lambda_j - 1\big| \leqslant \sum_{\ell\in[N]} \widetilde{\phi}_\ell + (r-1)(\lambda_1/\lambda_j) \prod_{\ell\in[N]} \widetilde{\phi}_\ell.$$

□

Theorem 4: Suppose Algorithm 3 (CPCA) is applied to the noiseless data $T^* = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk}$ through the SVD of $\text{mat}_S(T^*)$ for some nontrivial subset $S \subset [N]$. Let $\psi_0 = (\sqrt{2} + 4\lambda_1/\lambda_{\min,\pm})\delta$ with $\delta = \delta_S \vee \delta_{S^c}$, where $S^c = [N]\backslash S$. Then,

$$\begin{aligned}&|\widehat{\lambda}_j^{\text{cpca}} - \lambda_j| \leqslant \sqrt{2}\delta\lambda_1,\\ &\big(\|\widehat{a}_{jk}^{\text{cpca}}\widehat{a}_{jk}^{\text{cpca}\top} - a_{jk}a_{jk}^\top\|_{\text{S}}^2\big) \wedge (1/2) \leqslant \psi_0^2/2,\end{aligned} \tag{26}$$

for the resulting $\widehat{\lambda}_j^{\text{cpca}}$ and $\widehat{a}_{jk}^{\text{cpca}}$. Suppose further that for $\delta_{\max} = \max_{k\leqslant N}\delta_k$,

$$\begin{aligned}&3\max\big\{\delta_{\max}, (\sqrt{r}+1)\psi_0\big\} \leqslant 1,\\ &4\sqrt{r-1}(\lambda_1/\lambda_r)\psi_0^{N-2} \leqslant \rho < 1.\end{aligned} \tag{27}$$

Let $\gamma_N \in (2-2/N, 2)$ be the solution of $\gamma_N^N - 2\gamma_N^{N-1} + 1 = 0$, e.g. $\gamma_3 = 1.618$, $\gamma_4 = 1.839$. If the resulting $\widehat{a}_{jk}^{\text{cpca}}$ is used as the initialization of Algorithm 4 (ICO), then

$$\max_{j\leqslant r}\big(2 - 2\big|a_{jk}^\top \widehat{a}_{jk}^{(m)}\big|\big)^{1/2} \leqslant \psi_{m,k} = \psi_0\rho^{\gamma_N^{m\;\;1\;N\;\;k\;\;1}},$$

$$\max_{j\leqslant r}\big|\widehat{\lambda}_j^{(m)}/\lambda_j - 1\big| \leqslant \textstyle\sum_{k=1}^N \psi_{m,k} + \rho\psi_{m,N},$$

and $\max_{1\leqslant k\leqslant N}\psi_{m,k} \leqslant \epsilon$ within m iterations, where $m = \lceil N^{-1}\{1 + (\log\gamma_N)^{-1}\log(\log(\psi_0/\epsilon)/\log(1/\rho))\}\rceil$.

Proof of Theorem 4: By definition $\text{mat}_S(T^*) = A_S \Lambda A_{S^c}^\top = \sum_{j=1}^r \widehat{\lambda}_j^{\text{cpca}} \widehat{u}_j \widehat{v}_j^\top$, so that for the $U = (u_1, \ldots, u_r)$ and $V = (v_1, \ldots, v_r)$ in Proposition 5 we have $\|a_{jS}a_{jS}^\top - u_j u_j^\top\|_{\text{S}} \leqslant \delta_S$, $\|a_{jS^c}a_{jS^c}^\top - v_j v_j^\top\|_{\text{S}} \leqslant \delta_{S^c}$, $1 \leqslant j \leqslant r$, and $\|\text{mat}_S(T^*) - U\Lambda V^\top\|_{\text{S}} \leqslant \sqrt{2}\lambda_1\delta$. These and Proposition 3 yield (26) as in the proof of Theorem 1. Moreover, under (28) we have $\psi_0^2 < 1$, so that $2(1 - |a_{j\ell}^\top \widehat{a}_{j\ell}^{\text{cpca}}|) \leqslant 2\big\|\widehat{a}_{j\ell}^{\text{cpca}}\widehat{a}_{j\ell}^{\text{cpca}\top} - a_{j\ell}a_{j\ell}^\top\big\|_{\text{S}}^2 \leqslant \psi_0^2$. Define $\psi'_{m,k} = \max_{j\leqslant r}\big(2 - 2\big|a_{jk}^\top \widehat{a}_{jk}^{(m)}\big|\big)^{1/2}$ with $\widehat{a}_{jk}^{(0)} = \widehat{a}_{jk}^{\text{cpca}}$. By Proposition 6 and (28), $\psi'_{1,1} \leqslant \rho\psi_0$ and similar to the proof of Theorem 1, we have $\psi'_{m,k} \leqslant \rho^{n\;\;m\;\;1\;N\;\;k}\psi_0$ with $n_1 = 1$, $n_2 = 2, \ldots, n_N = 2^{N-1}$, and $n_{k+1} = 1 + \sum_{\ell=1}^{N-1} n_{k+1-\ell}$ for $k > N$. By induction, for $k = N, N+1, \ldots$.

$$n_{k+1} \geqslant \gamma_N^{k-1} + \cdots + \gamma_N^{k-N+1} = \gamma_N^k \frac{1-\gamma_N^{-N+1}}{\gamma_N - 1} = \gamma_N^k.$$

The function $f(\gamma) = \gamma^N - 2\gamma^{N-1} + 1$ is decreasing in $(1, 2-2/N)$ and increasing $(2-2/N, \infty)$. Because $f(1) = 0$ and $f(2) = 1 > 0$, we have $2 - 2/N < \gamma_N < 2$. By Proposition 6, (28) and the upper bound for $\psi'_{m,k}$, we have the desired upper bound for $\max_{j\leqslant r}\big|\widehat{\lambda}_j^{(m)}/\lambda_j - 1\big|$. □

Proposition 7: Let SP $:= \text{span}\{a_1, \ldots, a_r\}$, where $a_j = \text{vec}(\otimes_{k=1}^N a_{jk}) \in \mathbb{R}^d$. The elements of SP can be viewed as either length d vectors or $d_1 \times \cdots \times d_N$ tensors. Suppose $N > 2$ and $\delta_k < 1$ for every $k = 1, \ldots, N$ in (6), then every rank-1 tensor in SP is one of a_j's up to a scalar.

Proof of Proposition 7: Suppose M is a rank-1 tensor in SP$=\text{span}\{a_1, \ldots, a_r\}$, where $a_j = \text{vec}(\otimes_{k=1}^N a_{jk})$. Thus, there exist coefficients $\beta_j, j \leqslant r$, such that

$$M = \beta_1\text{vec}(\otimes_{k=1}^N a_{1k}) + \cdots + \beta_r\text{vec}(\otimes_{k=1}^N a_{rk}).$$

In matrix form, it follows that

$$\begin{aligned}&\text{mat}_1(M)\\ =&\beta_1 a_{11}\text{vec}(\otimes_{k=2}^N a_{1k})^\top + \cdots + \beta_r a_{r1}\text{vec}(\otimes_{k=2}^N a_{rk})^\top,\end{aligned}$$

where $\{a_{j1}, j \leqslant r\}$ is a set of linearly independent vectors, and $\{\text{vec}(\otimes_{k=2}^N a_{jk}), j \leqslant r\}$ is also a set of linearly independent vectors. Note that the matrix on the left hand side has rank 1 while the matrix on the right hand side has rank $|j \in [r] : \beta_j \neq 0|$. Since the rank of a matrix is unambiguously determined, we must have $|j \in [r] : \beta_j \neq 0| = 1$. Hence, $M = \beta_j\ \ a_j$ holds for some $j_* \in [r]$. □

APPENDIX B
ANALYSIS OF CPCA AND ICO FOR NOISY TENSORS

This section provides the analysis of the CPCA and ICO algorithms in the noisy case of models (2) and (4). In addition to the propositions provided before, we use concentration inequalities to derive the statistical error bounds.

Theorem 2: Consider spiked covariance tensor model (2) with data in (1), $\lambda_j = \omega_j^2$ and $\delta = \|A^\top A - I_r\|_{\text{S}}$ as in (8). In an event with probability at least $1 - e^{-t}$, Algorithm 1

(CPCA) gives the following error bound for the estimation of the CP basis vectors a_{jk},

$$\begin{aligned}&\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}} \\ &\leqslant (1+2\lambda_1/\lambda_{j,\pm})\delta + C(\lambda_1/\lambda_{j,\pm})\big(R^{(0)} + \sqrt{t/n}\big) \qquad (20)\end{aligned}$$

for all $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant K$ and $0 \leqslant t \leqslant d$, where C is a numeric constant, $\lambda_{j,\pm}$ is the j-th eigengap, and $R^{(0)} = \sqrt{(r_{\mathrm{eff}}/n)(1+1/\mathrm{SNR})(1+(r_{\mathrm{eff}}/d)/\mathrm{SNR})} \leqslant \sqrt{(r+\sigma^2 d/\lambda_1)(1+\sigma^2/\lambda_1)/n}$.

Proof of Theorem 2: Recall that $\lambda_j = w_j^2$ with $\lambda_1 \geqslant \cdots \geqslant \lambda_r > 0$, $A = (A_1,\ldots,A_r)$ with $a_j = \mathrm{vec}(a_{j1}\otimes a_{j2}\otimes\cdots\otimes a_{jK})$, $T = n^{-1}\sum_{i=1}^n \mathcal{X}_i \otimes \mathcal{X}_i$ and $d = d_1 d_2 \ldots d_K$. Write

$$\begin{aligned}\mathrm{mat}_{[K]}(T) &= \sum_{j=1}^r \lambda_j\big(\mathrm{vec}(\otimes_{k=1}^K a_{jk})\big)^{\otimes 2} + \sigma^2 I_d + \Psi^* \\ &= A\Lambda A^\top + \sigma^2 I_d + \Psi^*, \qquad (38)\end{aligned}$$

where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_r)$ and $\Psi^* = \mathrm{mat}_{[K]}(T - \mathbb{E}[T]) = \mathrm{mat}_{[K]}(\Psi) - \sigma^2 I_d$. Let $U = (u_1,\ldots,u_r)$ be the orthonormal matrix corresponding to A as in Proposition 2. We have $\|AA^\top - UU^\top\|_{\mathrm{S}} \leqslant \delta$ and $\|A\Lambda A^\top - U\Lambda U^\top\|_{\mathrm{S}} \leqslant \lambda_1\delta$ by two applications of the error bound in Proposition 2 with $\Lambda = I_r$ the first time. Let the top r eigenvectors of $\mathrm{mat}_{[K]}(T)$ be $\widehat{U} = (\widehat{u}_1,\ldots,\widehat{u}_r) \in \mathbb{R}^{d\times r}$. By Wedin's perturbation theorem [62] for any $1 \leqslant j \leqslant r$,

$$\begin{aligned}\|\widehat{u}_j\widehat{u}_j^\top - u_j u_j^\top\|_{\mathrm{S}} &\leqslant 2\|A\Lambda A^\top - U\Lambda U^\top + \Psi^*\|_{\mathrm{S}}/\lambda_{j,\pm} \\ &\leqslant \big(2\lambda_1\delta + 2\|\Psi^*\|_{\mathrm{S}}\big)/\lambda_{j,\pm}. \qquad (39)\end{aligned}$$

Combining (39) and the inequality $\|AA^\top - UU^\top\|_{\mathrm{S}} \leqslant \delta$, we have

$$\|\widehat{u}_j\widehat{u}_j^\top - a_j a_j^\top\|_{\mathrm{S}} \leqslant \delta + \big(2\lambda_1\delta + 2\|\Psi^*\|_{\mathrm{S}}\big)/\lambda_{j,\pm} \qquad (40)$$

We formulate each $\widehat{u}_j \in \mathbb{R}^d$ to be a K-way tensor $\widehat{U}_j \in \mathbb{R}^{d_1\times\cdots\times d_K}$. Let $\widehat{U}_{jk} = \mathrm{mat}_k(\widehat{U}_j)$, which is viewed as an estimate of $a_{jk}\mathrm{vec}(\otimes_{l\neq k}^K a_{jl})^\top \in \mathbb{R}^{d_k\times(d/d_k)}$. Then $\widehat{a}_{jk}^{\mathrm{cpca}}$ is the top left singular vector of $\widehat{U}_{jk}$. By Proposition 3,

$$\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}^2 \wedge (1/2) \leqslant \|\widehat{u}_j\widehat{u}_j^\top - a_j a_j^\top\|_{\mathrm{S}}^2. \quad (41)$$

Substituting (40) and Lemma 1 into the above equation, based on the definition of SNR and $R^{(0)}$, we have the desired results. We note that (20) holds automatically when the right-hand side is greater than 1, e.g. $\delta \geqslant 1$. □

Lemma 1: Suppose the assumptions in Theorem 2 hold and $\delta < 1$. Let $\Psi^* = \mathrm{mat}_{[K]}(T - \mathbb{E}[T])$ and $\lambda_j = w_j^2$ in (2). In an event with probability at least $1 - e^{-t}$, we have

$$\begin{aligned}&1 \wedge \big(\|\Psi^*\|_{\mathrm{S}}/\lambda_1\big) \\ &\leqslant C\max\Big(\sqrt{(r_{\mathrm{eff}}/n)(1+1/\mathrm{SNR})(1+(r_{\mathrm{eff}}/d)/\mathrm{SNR})}, \sqrt{t/n}\Big)\end{aligned}$$

for all $0 \leqslant t \leqslant d$, where C is a numerical constant.

Proof: Let $T^* = \mathbb{E}[T]$. As $\Psi^* = \mathrm{mat}_{[K]}(T - T^*)$, it follows from Theorem 2 of [63] that with probability at most e^{-t},

$$\|\Psi^*\|_{\mathrm{S}} \geqslant C\|\mathrm{mat}_{[K]}(T^*)\|_{\mathrm{S}}\big(\sqrt{(r^*\vee t)/n} \vee ((r^*\vee t)/n)\big)$$

where $r^* = \mathrm{trace}(\mathrm{mat}_{[K]}(T^*))/\|\mathrm{mat}_{[K]}(T^*)\|_{\mathrm{S}}$ is the effective rank of $\mathrm{mat}_{[K]}(T^*)$ and C is a numeric constant. Because $\mathrm{mat}_{[K]}(T^*) = \mathbb{E}[\mathrm{mat}_{[K]}(T)] = \sum_{j=1}^r \lambda_j a_j a_j^\top + \sigma^2 I_d$, $\|\mathrm{mat}_{[K]}(T^*)\|_{\mathrm{S}} \leqslant 2\lambda_1 + \sigma^2$ and $\mathrm{trace}(\mathrm{mat}_{[K]}(T^*)) = \lambda_1 r_{\mathrm{eff}} + \sigma^2 d = \lambda_1 r_{\mathrm{eff}}(1+1/\mathrm{SNR})$ with $r_{\mathrm{eff}} \leqslant r \leqslant d$, so that

$$\begin{aligned}&\min\left\{\frac{\|\mathrm{mat}_{[K]}(T^*)\|_{\mathrm{S}}\big(\sqrt{(r^*\vee t)/n} \vee ((r^*\vee t)/n)\big)}{3\lambda_1}, 1\right\} \\ &\leqslant \max\Big(\sqrt{(r_{\mathrm{eff}} + \sigma^2 d/\lambda_1)(2/3 + \sigma^2/(3\lambda_1))/n}, \\ &\qquad (2/3 + \sigma^2/(3\lambda_1))\sqrt{t/n}\Big) \\ &\leqslant \max\Big(\sqrt{(r_{\mathrm{eff}}/n)(1+1/\mathrm{SNR})(1+(r_{\mathrm{eff}}/d)/\mathrm{SNR})}, \sqrt{t/n}\Big).\end{aligned}$$

Note the component of the maximum with $\sqrt{t/n}$ is smaller when $\lambda_1 \leqslant \sigma^2$ and $0 \leqslant t \leqslant d$. □

Theorem 3: Suppose that with a proper numeric constant C_0 and the quantities defined in (21), (22) and (23),

$$\alpha > 0,\ \rho_1 \leqslant \rho < 1,\ C_{0,\alpha}R_{rK,1}^{(\mathrm{ideal})} \leqslant \psi_0 < 1. \qquad (24)$$

Let $\Omega_0 = \{\max_{j,k}\|\widehat{a}_{jk}^{(0)}\widehat{a}_{jk}^{(0)} - a_{jk}a_{jk}^\top\|_{\mathrm{S}} \leqslant \psi_0\}$ for any initial estimates $\widehat{a}_{jk}^{(0)}$. Then, Algorithm 2 (ICO) provides

$$\begin{aligned}&\mathbb{P}\left\{\max_{j,k}\min_{\Pi_r\in\mathscr{P}}\frac{\big\|\widehat{A}_k^{\mathrm{ico}}\Pi_r - A_k\big\|_{\mathrm{F}}}{(4r/3)^{1/2}(\epsilon_{rk}\vee\epsilon)} \leqslant 1\right\} \\ \geqslant\ &\mathbb{P}\left\{\max_{j,k}\frac{\|\widehat{a}_{jk}^{\mathrm{ico}}\widehat{a}_{jk}^{\mathrm{ico}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}}{\epsilon_{jk}\vee\epsilon} \leqslant 1\right\} \\ \geqslant\ &\mathbb{P}\big\{\Omega_0\big\} - mrKe^{-2(d_1\wedge\sqrt{n})} \qquad (25)\end{aligned}$$

within $m \geqslant m_\epsilon + 3$ iterations, where $\epsilon_{jk} = C_{0,\alpha}R_{jk,\phi_0}^{(\mathrm{ideal})}$, $m_\epsilon = \lceil\log(\log(\epsilon/\psi_0)/\log\rho)/\log 2\rceil$ for $(\epsilon_{r2}\vee\epsilon_0)\wedge\epsilon_{r3} \leqslant \epsilon < \psi_0$ and $m_\epsilon = \lceil\log(\epsilon/\psi_0)/\log\rho\rceil$ for $\epsilon_{r2} \leqslant \epsilon < \epsilon_0 \wedge \epsilon_{r3}$, with $\epsilon_0 = C_{0,\alpha}r/n$. Moreover, (25) holds within $m_{\epsilon_{r2}}+4$ iterations for $\epsilon = \epsilon_* \vee \sqrt{\epsilon_*\epsilon_0}$ where $\epsilon_* = C_{0,\alpha}(\lambda_1/\lambda_r)\prod_{k=2}^K \epsilon_{rk}^2$. In particular, if Algorithm 1 (CPCA) is used to initialize Algorithm 2 and ψ_0 is taken as the maximum of the right-hand side of (20), then (25) holds with $\mathbb{P}\{\Omega_0\} \geqslant 1 - e^{-t}$.

Proof of Theorem 3: We divide the proof into three steps.

Step 1 (Error Bound for a Single Update): Consider given (j,k) in this step. Recall that $(b_{1k},\ldots,b_{rk}) = A_k(A_k^\top A_k)^{-1}$ with $A_k = (a_{1k},\ldots,a_{rk})$. Let $z_n \sim N(0, I_n)$. For $g = \{g_1,\ldots,g_{2K}\}$ with $g_k = g_{k+K} \in \mathbb{R}^{d_k}$, define $T_k(g)$ as

$$T \times_{\ell\in[2K]\backslash\{k,k+K\}} g_\ell^\top = X_k(g)^\top X_k(g)/n \in \mathbb{R}^{d_k\times d_k}$$

with $X_k(g) = (\mathcal{X}_i \times_{\ell\in[K]\backslash\{k\}} g_\ell, i\in[n])^\top \in \mathbb{R}^{n\times d_k}$. Write

$$X_k(g) = M_{jk}(g) + M_{jk}^c(g) + E_k(g)$$

where $E_k(g) = \big(\mathcal{E}_i \times_{\ell\in[K]\backslash\{k\}} g_\ell^\top, i\in[n]\big)^\top \in \mathbb{R}^{n\times d_k}$,

$$M_{jk}(g) = \big(f_{ij}, i\in[n]\big)^\top\big(w_j\textstyle\prod_{\ell\in[K]\backslash\{k\}} a_{j\ell}^\top g_\ell\big)a_{jk}^\top$$

as a rank-one $n\times d_k$ random matrix with signal, and

$$M_{jk}^c(g) = \textstyle\sum_{h\in[r]\backslash\{j\}} M_{hk}(g) \in \mathbb{R}^{n\times d_k}.$$

As $T_k(g) = T \times_{\ell\in[2K]\backslash\{k,k+K\}} g_\ell^\top$, it follows that

$$T_k(g) = \bar{\lambda}_j(g)a_{jk}a_{jk}^\top + \sigma^2 I_{d_k} + \Delta_{jk}(g), \qquad (42)$$

where $\Delta_{jk}(g) = \sum_{i=1}^{5} \Delta_{jk}^{(i)}(g)$,

$$\begin{aligned}
\bar{\lambda}_j(g) &= \lambda_j \{\textstyle\prod_{\ell\in[K]\setminus\{k\}} (a_{j\ell}^\top g_\ell)^2\}\textstyle\sum_{i=1}^n f_{ij}^2/n,\\
\Delta_{jk}^{(1)}(g) &= M_{jk}^{c\top}(g)M_{jk}^c(g)/n,\\
\Delta_{jk}^{(2)}(g) &= E_{jk}^\top(g)E_{jk}(g)/n - \sigma^2 I_{d_k},\\
\Delta_{jk}^{(3)}(g) &= E_{jk}^\top(g)M_{jk}^c(g)/n + M_{jk}^{c\top}(g)E_{jk}(g)/n,\\
\Delta_{jk}^{(4)}(g) &= E_{jk}^\top(g)M_{jk}(g)/n + M_{jk}^\top(g)E_{jk}(g)/n,\\
\Delta_{jk}^{(5)}(g) &= M_{jk}^\top(g)M_{jk}^c(g)/n + M_{jk}^{c\top}(g)M_{jk}(g)/n.
\end{aligned}$$

We bound $\bar{\lambda}_j(g)$ and $\|\Delta_{jk}(g)\|_{\mathrm{S}}$ over $g_\ell \in G_{j\ell}$ with

$$\begin{aligned}
G_{j\ell} = \{g_\ell \in \mathbb{S}^{d_\ell-1} : \|g_\ell - b_{j\ell}/\|b_{j\ell}\|_2\|_2 \leqslant \phi, |a_{j\ell}^\top g_\ell| \geqslant \alpha,\\
\max\nolimits_{h\neq j}|a_{h\ell}^\top g_\ell| \leqslant \psi_\ell' = \psi_\ell/\sqrt{1-1/(4r)}\}
\end{aligned} \tag{43}$$

for $\ell \neq k$. In addition, we set $G_\ell = \mathbb{S}^{d_\ell-1}$ and $G_{jk} = G_k$.

By the Gaussian concentration of $(\sum_{i=1}^n f_{ij}^2)^{1/2}$,

$$\inf_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \bar{\lambda}_j(g) \geqslant \frac{\lambda_j \alpha^{2K-2}}{(1-1/\sqrt{n}-\sqrt{2t/n})^{-2}}$$

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

Similarly, in an event with at least probability $1-e^{-t}$,

$$\left\|\textstyle\sum_{i=1}^n F_i F_i^\top/n\right\|_{\mathrm{S}} \leqslant (1+\sqrt{r/n}+\sqrt{2t/n})^2$$

with $F_i = (f_{i1}, \ldots, f_{ir})^\top$, and in the same event

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \left\|\Delta_{jk}^{(1)}(g)\right\|_{\mathrm{S}}\\
\leqslant\ & \frac{\sup_{g_\ell\in G_{j\ell},\ell\in[K]}\sum_{h\neq j}\lambda_h \prod_{\ell\in[K]}(a_{h\ell}^\top g_\ell)^2}{(1+\sqrt{r/n}+\sqrt{2t/n})^{-2}}\\
=\ & \frac{\left\|\sum_{h\in[r]\setminus\{j\}}\lambda_h a_{hk}a_{hk}^\top\right\|_{\mathrm{S}}\left(\prod_{\ell\in[K]\setminus\{k\}}\psi_\ell'\right)^2}{(1+\sqrt{r/n}+\sqrt{2t/n})^{-2}}\\
\leqslant\ & \frac{\lambda_1(1+\delta_k)\left(\prod_{\ell\in[K]\setminus\{k\}}\psi_\ell'\right)^2}{(1+\sqrt{r/n}+\sqrt{2t/n})^{-2}}.
\end{aligned}$$

Let $\phi' = \phi \wedge 1$. For the noise component, the Sudakov-Fernique and Gaussian concentration inequalities provide

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \|E_k(g)\|_{\mathrm{S}}\\
\leqslant\ & \sigma\mathbb{E}[\|z_n\|_2] + \sigma\sqrt{2t} + \mathbb{E}\Big[\sup_{g_\ell\in G_{j\ell}\forall\ell\in[K]} \mathcal{E}_1 \times_{\ell=1}^K g_\ell\Big]\\
=\ & \sigma\Big(\mathbb{E}[\|z_n\|_2] + \sqrt{2t} + \mathbb{E}[\|z_{d_k}\|_2] + \phi'\sum_{\ell\neq k}\mathbb{E}[\|z_{d_\ell}\|_2]\Big)
\end{aligned}$$

with at least probability $1-e^{-t}$. Similarly, the smallest singular value $\sigma_1(E_k(g))$ is bounded from below by

$$\begin{aligned}
&\inf_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \sigma_1\big(E_k(g)\big)\\
\geqslant\ & \sigma\Big(\mathbb{E}[\|z_n\|_2] - \sqrt{2t} - \mathbb{E}[\|z_{d_k}\|_2] - \phi'\sum_{\ell\neq k}\mathbb{E}[\|z_{d_\ell}\|_2]\Big)
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \left\|\Delta_{jk}^{(2)}(g)\right\|_{\mathrm{S}}\\
\leqslant\ & \sigma^2\left\{\left(1+\frac{\sqrt{2t}+\sqrt{d_k}}{\sqrt{n}} + \phi'\sum_{\ell\neq k}\frac{\sqrt{d_\ell}}{\sqrt{n}}\right)^2 - 1\right\}
\end{aligned}$$

with at least probability $1-2e^{-t}$.

For each of the three cross-product terms, the two matrix factors are independent. Thus, an application of the above calculation in the proof of Lemma G.2 of [36] yields

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \left\|\Delta_{jk}^{(3)}(g)\right\|_{\mathrm{S}}\\
\leqslant\ & 2\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \|E_k^\top(g)M_{jk}^c(g)\|_{\mathrm{S}}/n\\
\leqslant\ & 2\sigma\sqrt{\lambda_1(1+\delta_k)}\Big(\prod_{\ell\in[K]\setminus\{k\}}\psi_\ell'\Big)\\
&\times\left\{\left(1+\frac{\sqrt{r}+\sqrt{2t}}{\sqrt{n}}\right)\left(\frac{\sqrt{d_k}}{\sqrt{n}}+\phi'\sum_{\ell\neq k}\frac{\sqrt{d_\ell}}{\sqrt{n}}\right)\right.\\
&\left.+\frac{\sqrt{r}}{\sqrt{n}}\left(1+\frac{\sqrt{2t}}{\sqrt{n}}\right)+\frac{2t+2\sqrt{2t}}{n}\right\}
\end{aligned}$$

with at least probability $1-2e^{-t}$,

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \left\|\Delta_{jk}^{(4)}(g)\right\|_{\mathrm{S}}\\
\leqslant\ & 2\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \|E_k^\top(g)M_{jk}(g)\|_{\mathrm{S}}/n\\
\leqslant\ & 2\sigma\lambda_j^{1/2}\left\{\left(1+\frac{1+\sqrt{2t}}{\sqrt{n}}\right)\left(\frac{\sqrt{d_k}}{\sqrt{n}}+\phi'\sum_{\ell\neq k}\frac{\sqrt{d_\ell}}{\sqrt{n}}\right)\right.\\
&\left.+\frac{1}{\sqrt{n}}+\frac{2t+3\sqrt{2t}}{n}\right\}
\end{aligned}$$

with at least probability $1-2e^{-t}$, and

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \left\|\Delta_{jk}^{(5)}(g)\right\|_{\mathrm{S}}\\
\leqslant\ & 2\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \|M_{jk}^\top(g)M_{jk}^c(g)\|_{\mathrm{S}}/n\\
\leqslant\ & 2\lambda_j^{1/2}\sqrt{\lambda_1(1+\delta_k)}\Big(\prod_{\ell\in[K]\setminus\{k\}}\psi_\ell'\Big)\\
&\times\left\{\left(1+\frac{1+\sqrt{2t}}{\sqrt{n}}\right)\frac{\sqrt{r}}{\sqrt{n}}+\frac{1}{\sqrt{n}}+\frac{2t+3\sqrt{2t}}{n}\right\}
\end{aligned}$$

with at least probability $1-2e^{-t}$.

Putting the above inequalities together, we find that for $r \leqslant n$ and with at least probability $1-e^{-2(d_1\wedge\sqrt{n})}$,

$$\inf_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \bar{\lambda}_j(g) \geqslant \lambda_j\alpha^{2K-2}/C_0' \tag{44}$$

and with $\psi_{-k} = \prod_{\ell\in[K]\setminus\{k\}}\psi_\ell$

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \left\|\Delta_{jk}(g)\right\|_{\mathrm{S}}\\
\leqslant\ & C_0'\lambda_1\psi_{-k}^2 + C_0'\sigma^2\big(d_{k,\phi}^{1/2}/n^{1/2} + d_{k,\phi}/n\big)\\
&+C_0'\lambda_j^{1/2}\sigma d_{k,\phi}^{1/2}/n^{1/2} + C_0'(\lambda_1\lambda_j r/n)^{1/2}\psi_{-k},
\end{aligned} \tag{45}$$

where $d_{k,\phi} = \big(d_k^{1/2} + (\phi\wedge 1)\sum_{\ell\in[K]\setminus\{k\}} d_\ell^{1/2}\big)^2$ and C_0' is a numeric constant. Here the upper bound for $\Delta_{jk}^{(3)}$ is absorbed into those for $\Delta_{jk}^{(1)}$ and $\Delta_{jk}^{(2)}$ by Cauchy-Schwarz.

Let $\widehat{a}_{jk}(g)$ be the top eigenvector of $T_k(g)$ in (42). As $\|a_{jk}\|_2 = \|\widehat{a}_{jk}(g)\|_2 = 1$, (42), (44) and (45) imply

$$\begin{aligned}
&\sup_{g_\ell\in G_{j\ell},\ell\in[K]\setminus\{k\}} \|\widehat{a}_{jk}(g)\widehat{a}_{jk}^\top(g) - a_{jk}a_{jk}^\top\|_{\mathrm{S}}\\
\leqslant\ & C_{0,\alpha}\max\Big\{(\lambda_1/\lambda_j)\psi_{-k}^2, R_{jk,\phi}^{(\mathrm{ideal})}, \sqrt{(\lambda_1/\lambda_j)(r/n)}\psi_{-k}\Big\}
\end{aligned} \tag{46}$$

with at least probability $1 - e^{-2(d_1 \wedge \sqrt{n})}$, where $G_{j\ell}$ are as in (43), $R^{(\text{ideal})}_{jk,\phi}$ as in (22) and $C_{0,\alpha} = C_0\alpha^{2-2K}$ with a numeric constant C_0. Here we assume C_0 can be taken as the constant in (23) and (24).

Step 2 (Error Bound Sequences): Recall that $\widehat{A}^{(m)}_\ell = (\widehat{a}^{(m)}_{1\ell}, \ldots, \widehat{a}^{(m)}_{r\ell}) \in \mathbb{R}^{d_\ell \times r}$, $\widehat{\Sigma}^{(m)}_\ell = \widehat{A}^{(m)\top}_\ell \widehat{A}^{(m)}_\ell$, and $\widehat{B}^{(m)}_\ell = \widehat{A}^{(m)}_\ell (\widehat{\Sigma}^{(m)}_\ell)^{-1} = (\widehat{b}^{(m)}_{1\ell}, \ldots, \widehat{b}^{(m)}_{r\ell}) \in \mathbb{R}^{d_\ell \times r}$. Let

$$\Omega_{m,\ell} = \Big\{ \max_{h \leqslant r} \|\widehat{a}^{(m)}_{h\ell}\widehat{a}^{(m)\top}_{h\ell} - a_{h\ell}a^\top_{h\ell}\|_{\mathrm{S}} \leqslant \psi_{m,\ell} \Big\} \tag{47}$$

with constants $\psi_{m,\ell} \leqslant \psi_0$ to be specified later sequentially. As the PCA of $T(g)$ in (42) does not depend on the signs of $g_{h\ell}$, we may assume without loss of generality $a^\top_{h\ell}\widehat{a}^{(m)}_{h\ell} \geqslant 0$ for all (h, ℓ). Thus, in $\Omega_{m,\ell}$ the proof of Proposition 4 provides

$$\max\nolimits_{h \leqslant r} \|\widehat{a}^{(m)}_{h\ell} - a_{h\ell}\|_2 \leqslant \psi_{m,\ell}/\sqrt{1 - 1/(4r)}, \tag{48}$$

$$\big\|\widehat{b}^{(m)}_{h\ell}\big\|_2 \leqslant \|\widehat{B}^{(m)}_\ell\|^{1/2}_{\mathrm{S}} \leqslant \left(\sqrt{1-\delta_\ell} - \frac{r^{1/2}\psi_0}{\sqrt{1-1/(4r)}}\right)^{-1}.$$

Let $P_\ell = A_\ell(A^\top_\ell A_\ell)^{-1}A^\top_\ell$ and $P^\perp_\ell = I_{d_\ell} - P^\top_\ell$. As $\widehat{B}^{(m)}_\ell - B_\ell = P^\perp_\ell\big(\widehat{A}^{(m)}_\ell - A_\ell\big)(\widehat{\Sigma}^{(m)}_\ell)^{-1} - B_\ell\big(\widehat{A}^{(m)}_\ell - A_\ell\big)^\top \widehat{B}^{(m)}_\ell$,

$$\begin{aligned} &\big\|\widehat{b}^{(m)}_{h\ell} - b_{h\ell}\big\|^2_2 \\ \leqslant\ & \big\|\widehat{A}^{(m)}_\ell - A_\ell\big\|^2_{\mathrm{S}}\big(\|\widehat{B}^{(m)}_\ell\|^2_{\mathrm{S}} + \|B_\ell\|^2_{\mathrm{S}}\big)\big(\|\widehat{b}^{(m)}_{h\ell}\|^2_2 \wedge \|b_{h\ell}\|^2_2\big) \\ \leqslant\ & \{r\psi^2_{m,\ell}/(1-1/(4r))\}(2/\alpha^2)\big(\|\widehat{b}^{(m)}_{h\ell}\|^2_2 \wedge \|b_{h\ell}\|^2_2\big) \end{aligned}$$

by the algebraic symmetry between the estimator and estimand, where α is as in (23). Let $\widehat{g}^{(m)}_{h\ell} = \widehat{b}^{(m)}_{h\ell}/\|\widehat{b}^{(m)}_{h\ell}\|_2$. As $\big\|\widehat{g}^{(m)}_{h\ell} - b_{h\ell}\big\|_2 \leqslant \big\|\widehat{b}^{(m)}_{h\ell} - b_{h\ell}\big\|_2$ for $\|\widehat{b}^{(m)}_{h\ell}\|_2 \geqslant \|b_{h\ell}\|_2 = 1$,

$$\big\|\widehat{g}^{(m)}_{h\ell} - b_{h\ell}/\|b_{h\ell}\|_2\big\|_2 \leqslant (\psi_{m,\ell}/\alpha)\sqrt{2r/(1-1/(4r))} \tag{49}$$

by scale invariance. Moreover, (48) provides

$$\max_{h \neq j}\big|a^\top_{h\ell}\widehat{g}^{(m)}_{j\ell}\big| \leqslant \psi_{m,\ell}/\sqrt{1-1/(4r)},\ \big|a^\top_{j\ell}\widehat{g}^{(m)}_{j\ell}\big| \geqslant \alpha, \tag{50}$$

as $\widehat{a}^{(m)\top}_{h\ell}\widehat{g}^{(m)}_{j\ell} = I\{h = j\}/\|\widehat{b}^{(m)}_{j\ell}\|_2$. Thus, in the event $\Omega_{m,\ell}$, $\widehat{g}^{(m)}_{h\ell} = \widehat{b}^{(m)}_{h\ell}/\|\widehat{b}^{(m)}_{h\ell}\|_2 \in G_{j,\ell}$ for $\ell \neq k$ in (43) with $\psi_\ell = \psi_{m,\ell}$, the α in (23) and any upper bound ϕ for (49).

Let

$$\Omega_{m,j,k} = \big\{\|\widehat{a}^{(m)}_{jk}\widehat{a}^{(m)\top}_{jk} - a_{jk}a^\top_{jk}\|_{\mathrm{S}} \leqslant \psi_{m,j,k}\big\}.$$

Let $\psi_{0,j,k} = \psi_{0,k} = \psi_0$ and sequentially update them by

$$\begin{aligned} \psi_{m,j,k} &= C_{0,\alpha}\Big\{\Big((\lambda_1/\lambda_j)\textstyle\prod_{\ell=1}^{K-1}\psi^2_{m,k-\ell}\Big) \vee R^{(\text{ideal})}_{jk,\phi_{m,k}} \\ &\qquad \vee \Big(\sqrt{r/n}\sqrt{\lambda_1/\lambda_j}\textstyle\prod_{\ell=1}^{K-1}\psi_{m,k-\ell}\Big)\Big\}, \\ \psi_{m,k} &= \psi_{m,r,k}, \end{aligned} \tag{51}$$

$k = 1, \ldots, K$, $m = 1, 2, \ldots$, with the $C_{0,\alpha}$ in (46) and

$$\phi_{m,k} = 1 \wedge \big(\max\nolimits_{1 \leqslant \ell < K}(\psi_{m,k-\ell}/\alpha)\sqrt{2r/(1-1/(4r))}\big).$$

Here and in the sequel, we take the convention $(m, \ell) = (m-1, K+\ell)$ with the subscript (m, ℓ). We note that $\psi_{m,k}$ depends on $\psi_{m,k-1}, \ldots, \psi_{m,k-K+1}$ only as an increasing function of their product and maximum. Thus, as $\psi_{1,k} \leqslant \psi_{0,k} = \psi_0$ by (24), $\psi_{m,k} \leqslant \psi_{m-1,k}$ for all $k \in [K]$ and $m \geqslant 1$ by induction.

By (46), (43), (49) and (50), the events $\Omega_{m,\ell}$ in (47) satisfy

$$\mathbb{P}\big\{\big(\cap_{\ell=1}^{K-1}\Omega_{m,k-\ell}\big) \cap \Omega^c_{m,j,k}\big\} \leqslant e^{-2(d_1 \wedge \sqrt{n})} \tag{52}$$

with $\cap_{j=1}^r \Omega_{m,j,k} \subseteq \Omega_{m,k}$.

Let $\phi_0 = (\psi^*/\alpha)\sqrt{2r/(1-1/(4r))}$ be as in (23) with $\psi^* = C_{0,\alpha}R^{(\text{ideal})}_{rK,1}$. A simple way of dealing with the dynamics of (51) is to compare $\psi_{m,j,k}$ with

$$\begin{aligned} \psi^*_{m,j,k} &= C_{0,\alpha}\Big\{\Big((\lambda_1/\lambda_j)\textstyle\prod_{\ell \neq k}\psi^{*2}_{m-1,\ell}\Big) \vee R^{(\text{ideal})}_{jk,\phi_0} \\ &\qquad \vee \Big(\sqrt{r/n}\sqrt{\lambda_1/\lambda_j}\textstyle\prod_{\ell \neq k}\psi^*_{m-1,\ell}\Big)\Big\}, \\ \psi^*_{m,k} &= \psi^*_{m,r,k}, \end{aligned} \tag{53}$$

with initialization $\psi^*_{0,j,k} = \psi_0$. Compared with (51), (53) is easier to analyze due to the use of static ϕ_0 and the monotonicity of $\psi^*_{m,k}$ in k. While (51) uses inputs with indices $(m, k - [K-1])$, (53) uses inputs with indices $(m-1, [K]\backslash\{k\})$. Thus, as $\max_{j,k,\phi} C_{0,\alpha}R^{(\text{ideal})}_{jk,\phi} \leqslant \psi^*$, $\psi_{m,k} \leqslant \psi^*_{m,k}$ before $\psi^*_{m,k}$ first hits $(0, \psi^*]$ at a certain (m^*, k^*). As $\psi^*_{m\ \ ,k} \leqslant \psi^*$ for $k \in [K]$, $\psi_{m\ \ ,k} \leqslant \psi^*$ for $k \in [K]$, so that $\phi_{m,k} \leqslant \phi_0$ for $m > m^*$. It follows that

$$\psi_{m+1,j,k} \leqslant \psi^*_{m,j,k} \leqslant \psi^*_{m,k}, \quad \forall (m, j, k). \tag{54}$$

Step 3 (Contraction of Error Bounds): Recall that $\epsilon_{jk} = C_{0,\alpha}R^{(\text{ideal})}_{jk,\phi_0}$. By (52) and (54), (25) follows from

$$\psi^*_{m,j,k} \leqslant \epsilon \vee \epsilon_{jk} \quad \forall j, k,$$

for $\epsilon_{r2} \leqslant \epsilon \leqslant \psi_0$ and $m \geqslant m_\epsilon + 2$. Let $\epsilon_{r,K+1} = \psi_0$. By induction, it suffices to prove that for $\epsilon_{rk_0} \leqslant \epsilon < \epsilon_{r,k_0+1}$

$$\psi^*_{m,j,k} \leqslant \epsilon \vee \epsilon'_{jk} \quad \forall m \geqslant m_\epsilon + 2, \forall j, k \tag{55}$$

where $\epsilon'_{rk} = \psi_0$ for $k > k_0$, $\epsilon'_{jk} = \epsilon_{jk}$ for $j < r$ or $k \leqslant k_0$, with each fixed $k_0 \geqslant 2$. This is done by comparing (53) with

$$\begin{aligned} \psi'_{m,j,k} &= \Big(C_{0,\alpha}(\lambda_1/\lambda_j)\textstyle\prod_{\ell \neq k}\psi'^2_{m-1,\ell}\Big) \vee \epsilon'_{jk} \\ &\qquad \vee \Big(C_{0,\alpha}\sqrt{r/n}\sqrt{\lambda_1/\lambda_j}\textstyle\prod_{\ell \neq k}\psi'_{m-1,\ell}\Big), \\ \psi'_{m,k} &= \psi'_{m,r,k}, \end{aligned} \tag{56}$$

with $\psi'_{0,j,k} = \psi_0$. Because $\epsilon_{jk} \leqslant \epsilon'_{jk}$,

$$\psi^*_{m,j,k} \leqslant \psi'_{m,j,k},\ \forall\ m, j, k. \tag{57}$$

Let $m_* = \min\{m : \psi'_{m,k_0} \leqslant \epsilon\}$. For $m < m_*$, $\psi'_{m,k} = \psi'_{m,k_0}$ for $k \leqslant k_0$ and $\psi'_{m,k} = \psi_0$ for $k_0 < k \leqslant K$, so that

$$\begin{gathered} C_{0,\alpha}(\lambda_1/\lambda_r)\psi'^{2(k_0-1)}_{m\ \ -1,k_0}\psi_0^{2(K-k_0)} \leqslant \epsilon, \\ C_{0,\alpha}\sqrt{r/n}\sqrt{\lambda_1/\lambda_r}\psi'^{k_0-1}_{m\ \ -1,k_0}\psi_0^{K-k_0} \leqslant \epsilon, \end{gathered}$$

which implies $\psi'_{m\ \ ,j,k} \leqslant \epsilon \vee \epsilon'_{jk}\ \forall j, k$, due to $\lambda_j \geqslant \lambda_r$ and $\psi'_{m\ \ -1,k_0} \leqslant \psi_0$. Consequently, (55) holds by (57). We note that when $\psi^*_{m,2} = \epsilon_{r2}$, $\psi^*_{m+1,j,k} \leqslant \epsilon_* \vee \sqrt{\epsilon_*\epsilon_0} \vee \epsilon_{jk}$.

It remains to prove $m_* \leqslant m_\epsilon + 2$. For $\epsilon \geqslant C_{0,\alpha}r/n$,

$$\epsilon < \psi'_{m,k_0} = \psi_0\rho^{1+(2k_0-2)+\cdots+(2k_0-2)^{m\ \ 1}}$$

with the $\rho < 1$ in (24) for $m < m_*$, so that

$$m_* - 2 \leqslant \lceil\log(\log(\epsilon/\psi_0)/\log\rho)/\log(2k_0-2)\rceil \leqslant m_\epsilon.$$

Let $m_1 = \min\{m : \psi'_{m,k_0} \leqslant C_{0,\alpha}r/n\}$ and $n_1 = \{1 + \ldots + (2k_0-2)^{m_1-2}\}I\{m_1 \geqslant 2\}$. For $\epsilon < C_{0,\alpha}r/n$, we have

$$\begin{aligned}\epsilon &< \psi'_{m,k_0}\\ &= \psi_0\rho_1^{1+\ldots+(k_0-1)^{m\ \ m_1}}\rho^{n_1(k_0-1)^{m\ \ m_1\ \ 1}}\\ &\leqslant \psi_0\rho^{1+\ldots+(k_0-1)^{m\ \ 1}}\end{aligned}$$

for $m_1 \leqslant m < m_*$, so that

$$\begin{aligned}&m_*-2\\ \leqslant\ &\begin{cases}\lceil\log(\log(\epsilon/\psi_0)/\log\rho)/\log(k_0-1)\rceil, & k_0>2,\\ \lceil\log(\epsilon/\psi_0)/\log\rho)\rceil, & k_0=2.\end{cases}\end{aligned}$$

Again $m_* \leqslant m_\epsilon + 2$. □

Theorem 5: Let $T = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk} + \Psi$ as in (4). Suppose $\Psi \in \mathbb{R}^{d_1\times\cdots\times d_N}$ has i.i.d $N(0,\sigma^2)$ entries. Then, in an event with probability at least $1-e^{-2d_S-2(d/d_S)}$, Algorithm 3 (CPCA) gives the following bound in the estimation of the CP basis vectors a_{jk}, $1 \leqslant j \leqslant r$, $1 \leqslant k \leqslant N$,

$$\begin{aligned}&\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}\\ \leqslant&(1+2\sqrt{2}\lambda_1/\lambda_{j,\pm})\delta + 6\sigma(\sqrt{d_S}+\sqrt{d/d_S})/\lambda_{j,\pm}\end{aligned} \tag{29}$$

where $\delta = \|A_S^\top A_S - I\|_{\mathrm{S}} \vee \|A_{S^c}^\top A_{S^c} - I_r\|_{\mathrm{S}}$ as in Theorem 4 and $\lambda_{j,\pm} = \min(\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1})$ are the eigengaps with $\lambda_0 = 2\lambda_1$ and $\lambda_{r+1} = 0$.

Proof of Theorem 5: Let $a_{j,S} = \mathrm{vec}(\otimes_{k\in S}a_{jk})$ and $a_{j,S^c} = \mathrm{vec}(\otimes_{k\in[N]\setminus S}\ a_{j,k})$. Let $U = (u_1,\ldots,u_r)$ and $V = (v_1,\ldots,v_r)$ be the orthonormal matrices in Proposition 5 with A and B there replaced respectively by A_S and A_{S^c}. By Proposition 5,

$$\begin{aligned}&\|a_{j,S}a_{j,S}^\top - u_ju_j^\top\|_{\mathrm{S}} \vee \|a_{j,S^c}a_{j,S^c}^\top - v_jv_j^\top\|_{\mathrm{S}} \leqslant \delta,\\ &\big\|\mathrm{mat}_S(T^*) - U\Lambda V^\top\big\|_{\mathrm{S}} \leqslant \sqrt{2}\delta\lambda_1,\end{aligned} \tag{58}$$

where $T^* = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk}$. Let $\Psi^* = \mathrm{mat}_S(\Psi) = \mathrm{mat}_S(T-T^*)$. We have

$$\big\|\mathrm{mat}_S(T) - U\Lambda V^\top\big\|_{\mathrm{S}} \leqslant \sqrt{2}\delta\lambda_1 + \|\Psi^*\|_{\mathrm{S}}.$$

As $\lambda_1 > \lambda_2 > \ldots > \lambda_r > \lambda_{r+1} = 0$, Wedin's perturbation theorem [62] provides

$$\begin{aligned}&\max\big\{\|\widehat{a}_{j,S}\widehat{a}_{j,S}^\top - u_ju_j^\top\|_{\mathrm{S}}, \|\widehat{a}_{j,S^c}\widehat{a}_{j,S^c}^\top - v_jv_j^\top\|_{\mathrm{S}}\big\}\\ \leqslant&\frac{2\sqrt{2}\lambda_1\delta + 2\|\Psi^*\|_{\mathrm{S}}}{\min\{\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1}\}}.\end{aligned} \tag{59}$$

Combining (58) and (59), we have

$$\begin{aligned}&\max\big\{\|\widehat{a}_{j,S}\widehat{a}_{j,S}^\top - a_{j,S}a_{j,S}^\top\|_{\mathrm{S}}, \|\widehat{a}_{j,S^c}\widehat{a}_{j,S^c}^\top - a_{j,S^c}a_{j,S^c}^\top\|_{\mathrm{S}}\big\}\\ \leqslant&\delta + \frac{2\sqrt{2}\lambda_1\delta + 2\|\Psi^*\|_{\mathrm{S}}}{\min\{\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1}\}}.\end{aligned} \tag{60}$$

By Theorem II.13 in [64], for any $x > 0$,

$$\mathbb{P}\left(\|\Psi^*\|_{\mathrm{S}}/\sigma > \sqrt{\prod_{k\in S} d_k} + \sqrt{\prod_{k\in[N]\setminus S} d_k} + x\right) \leqslant e^{-x^2/2}.$$

It implies that, choosing $x = 2\sqrt{d_S} + 2\sqrt{d_{S^c}}$, in an event with probability at least $1-e^{-2d_S-2d_{S^c}}$,

$$\|\Psi^*\|_{\mathrm{S}} \leqslant 3\sigma\sqrt{d_S} + 3\sigma\sqrt{d_{S^c}}. \tag{61}$$

We formulate each $\widehat{u}_j \in \mathbb{R}^d$ to be a K-way tensor $\widehat{U}_j \in \mathbb{R}^{d_1\times\cdots\times d_K}$. Let $\widehat{U}_{jk} = \mathrm{mat}_k(\widehat{U}_j)$, which is viewed as an estimate of $a_{jk}\mathrm{vec}(\otimes_{l\in S\setminus\{k\}}\ a_{jl})^\top \in \mathbb{R}^{d_k\times(d_S/d_k)}$. Then $\widehat{a}_{jk}^{\mathrm{cpca}}$ is the top left singular vector of $\widehat{U}_{jk}$. By Proposition 3, for any $k \in S$

$$\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}^2 \wedge (1/2) \leqslant \|\widehat{a}_{j,S}\widehat{a}_{j,S}^\top - a_{j,S}a_{j,S}^\top\|_{\mathrm{S}}^2.$$

Similar bound can be obtained for $\|\widehat{a}_{jk}^{\mathrm{cpca}}\widehat{a}_{jk}^{\mathrm{cpca}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}}$ for $k \in S^c$. Substituting (60) and (61) into the above equation, we have the desired results. □

Theorem 6: Let data T be as in Theorem 5 and $\Omega_0 = \{\max_{j,k}(2-2|a_{jk}^\top\widehat{a}_{jk}^{(0)}|)^{1/2} \leqslant \psi_0\}$ for any initial estimates $\widehat{a}_{jk}^{(0)}$. Let $\mathscr{P}_\pm$ be as in (25). Suppose

$$\alpha_* > 0,\ \rho^* < 1,\ 6\alpha_*^{1-N}R_{rK,1}^{*(\mathrm{ideal})} \leqslant \psi_0 < 1, \tag{33}$$

with the quantities defined in (30), (31) and (32). Then, in an event with probability at least $\mathbb{P}\{\Omega_0\} - e^{-d_N} - \sum_{k=1}^N e^{-d_k}$, Algorithm 4 (ICO) provides

$$|\widehat{\lambda}_j^{\mathrm{ico}}/\lambda_j - 1| \leqslant \epsilon_{jN}^* \vee \epsilon, \tag{34}$$

$$\|\widehat{a}_{jk}^{\mathrm{ico}}\widehat{a}_{jk}^{\mathrm{ico}\top} - a_{jk}a_{jk}^\top\|_{\mathrm{S}} \leqslant \epsilon_{jk}^* \vee \epsilon, \tag{35}$$

$$\min_{\Pi_r\in\mathscr{P}}\ \|\widehat{A}_k^{\mathrm{ico}}\Pi_r - A_k\|_{\mathrm{F}} \leqslant r^{1/2}(\epsilon_{rk}^* \vee \epsilon), \tag{36}$$

simultaneously for all $1 \leqslant j \leqslant r$ and $1 \leqslant k \leqslant N$, within $m \geqslant m_\epsilon + 3$ iterations, where $\epsilon_{jk}^* = 6\alpha_*^{N-1}R_{jk,\phi_0}^{*(\mathrm{ideal})}$ and $m_\epsilon = \lceil\log(\log(\epsilon/\psi_0)/\log\rho^*)/\log 2\rceil$ for $\epsilon_{r2}^* \leqslant \epsilon < \psi_0$. Moreover, (34), (35) and (36) hold in the same event within $m_{\epsilon_{r2}} + 4$ iterations for $\epsilon = 6\alpha_*^{1-N}\sqrt{r-1}(\lambda_1/\lambda_r)\prod_{k=2}^N \epsilon_{rk}^*$. If Algorithm 3 (CPCA) is used as initialization, then $\mathbb{P}\{\Omega_0\} \geqslant 1-\sum_{k=1}^N e^{-2d_k}$ for $\psi_0 = 6[\lambda_1\delta + \sigma(\sqrt{d_S}+\sqrt{d/d_S})]/\lambda_{\min,\pm}$.

Proof of Theorem 6: Let $\psi_{0,\ell} = \psi_0$ and define sequentially

$$\begin{aligned}\phi_{m,k-1}^* &= (N-1)\alpha_*^{-1}\sqrt{2r}\max_{1\leqslant\ell<N}\psi_{m,k-\ell},\\ \psi_{m,k} &= \Big(6\alpha_*^{1-N}\sqrt{r-1}(\lambda_1/\lambda_j)\textstyle\prod_{\ell=1}^{N-1}\psi_{m,k-\ell}\Big)\\ &\quad\vee\Big(6\alpha_*^{1-N}R_{jk,\phi_{m,k\ \ 1}}^{*(\mathrm{ideal})}\Big),\end{aligned} \tag{62}$$

$k = 1,\ldots,N$, $m = 1,2,\ldots$ By induction, (33) gives $\psi_{m,k} \leqslant \psi_{m-1,k} \leqslant \psi_0$. Here and in the sequel, we take the convention that $(m,\ell) = (m-1, N+\ell)$ with the subscript (m,ℓ), and that $\times_\ell\widehat{\theta}_{j,\ell}^{(m)} = \times_{N+\ell}\widehat{\theta}_{j,N+\ell}^{(m-1)}$ for any estimator $\widehat{\theta}_{jk}^{(m)}$. Let

$$\Omega_{m,k-1}^* = \cap_{\ell=1}^{N-1}\Omega_{m,k-\ell} \tag{63}$$

with $\Omega_{m,\ell} = \big\{\max_{h\leqslant r}\big(2-2|a_{h\ell}^\top\widehat{a}_{h\ell}^{(m)}|\big)^{1/2} \leqslant \psi_{m,\ell}\big\}$. Let $\widehat{g}_{j\ell}^{(m)} = \widehat{b}_{j\ell}^{(m)}/\|\widehat{b}_{j\ell}^{(m)}\|_2$, and $g_{j\ell} = b_{j\ell}/\|b_{j\ell}\|_2$. By (48), (49) and (50) in the proof of Theorem 3,

$$\big\|\widehat{g}_{j\ell}^{(m)} - g_{j\ell}\big\|_2 \leqslant (\psi_{m,\ell}/\alpha_*)\sqrt{2r},\quad \big|a_{j\ell}^\top\widehat{g}_{j\ell}^{(m)}\big| \geqslant \alpha_*, \tag{64}$$

in $\Omega_{m,\ell}$ with the α_* in (32).

Given $\{\widehat{a}_{j,k-\ell}^{(m)}, j \in [r], \ell \in [N-1]\}$, the m-th iteration for tensor mode k produces estimates $\widehat{a}_{jk}^{(m)}$ as the normalized version of $T \times_{\ell=k-1}^{k-N+1}\widehat{b}_{j,\ell}^{(m)\top}$. Because $T = \sum_{j=1}^r \lambda_j \otimes_{k=1}^N a_{jk} + \Psi$, the "noiseless" version of this update is given by

$$T \times_{\ell\in[N]\setminus\{k\}} b_{j\ell}^\top = \lambda_j a_{jk} + \Psi \times_{\ell\in[N]\setminus\{k\}} b_{j\ell}^\top \in \mathbb{R}^{d_k}. \tag{65}$$

Similarly, for any $1 \leqslant j \leqslant r$,

$$T \times_{\ell=k-1}^{k-N+1} \widehat{b}_{j,\ell}^{(m)\top} = \sum_{h=1}^{r} \widetilde{\lambda}_{h,j} a_{hk} + \Psi \times_{\ell=k-1}^{k-N+1} \widehat{b}_{j,\ell}^{(m)\top} \in \mathbb{R}^{d_k},$$

where $\widetilde{\lambda}_{h,j} = \lambda_h \prod_{\ell=1}^{N-1} a_{h,k-\ell}^\top \widehat{b}_{j,k-\ell}^{(m)}$. Let

$$\widetilde{\phi}_{m,\ell} = \psi_{m,\ell} / \big(\sqrt{1-\delta_{\max}} - \sqrt{r}\psi_{m,\ell}\big)_+.$$

By the definition of α_* in (32) and the condition $\psi_{m,\ell} \leqslant \psi_0$, $\widetilde{\phi}_{m,\ell}/(1-\widetilde{\phi}_{m,\ell}) \leqslant \psi_{m,\ell}/\alpha_*$. Thus, by the arguments in the proof of Proposition 6,

$$\begin{aligned} &\big(2 - 2|a_{jk}^\top \widehat{a}_{jk}^{(m)}|\big)^{1/2} \\ \leqslant &\frac{\sqrt{2}\big\|\Psi \times_{\ell=k-1}^{k-N+1} \widehat{g}_{j,\ell}^{(m)\top}\big\|_2}{\lambda_j \prod_{\ell=1}^{N-1} a_{j,k-\ell}^\top \widehat{g}_{j,k-\ell}^{(m)}} + \frac{\lambda_1\sqrt{2(1+\delta_k)}}{\lambda_j/\sqrt{r-1}} \prod_{\ell=1}^{N-1} \frac{\psi_{m,k-\ell}}{\alpha_*} \end{aligned} \tag{66}$$

in $\Omega^*_{m,k-1}$. As $\Psi \times_{\ell=k-1}^{k-N+1} \widehat{g}_{j,\ell}^{(m)\top}$ is linear in each $\widehat{g}_{j\ell}^{(m)}$,

$$\begin{aligned} &\big\|\Psi \times_{\ell=k-1}^{k-N+1} \widehat{g}_{j,\ell}^{(m)\top}\big\|_2 \\ \leqslant\ & (N-1)\max_{\ell<N} \|\widehat{g}_{j,k-\ell}^{(m)} - g_{j,k-\ell}\|_2 \|\Delta\| \\ & + \big\|\Psi \times_{\ell \in [N]\setminus\{k\}} g_{j\ell}^\top\big\|_2, \end{aligned}$$

where $\|\Delta\| = \max_{v_\ell \in \mathbb{S}^{d_\ell - 1} \forall \ell} \big(\Psi \times_{\ell=1}^{N} v_\ell^\top\big)$. As we also have $\big\|\Psi \times_{\ell=k-1}^{k-N+1} \widehat{g}_{j,\ell}^{(m)\top}\big\|_2 \leqslant \|\Delta\|$, (64) yields

$$\begin{aligned} &\big\|\Psi \times_{\ell=k-1}^{k-N+1} \widehat{g}_{j,\ell}^{(m)\top}\big\|_2 \\ \leqslant\ & \min\big\{\|\Delta\|, \phi^*_{m,k-1}\|\Delta\| + \big\|\Psi \times_{\ell\in[N]\setminus\{k\}} g_{j\ell}^\top\big\|_2\big\} \end{aligned} \tag{67}$$

in $\Omega^*_{m,k-1}$, in view of the definition of $\phi^*_{m,k-1}$ in (62). By the Sudakov-Fernique and Gaussian concentration inequalities,

$$\mathbb{P}\left(\|\Delta\|/\sigma > \sum_{\ell=1}^{N} \sqrt{d_\ell} + x\right) \leqslant e^{-x^2/2}$$

and $\mathbb{P}\{\|\Psi \times_{\ell\in[N]\setminus\{k\}} g_{j\ell}^\top\|_2 > \sqrt{d_k} + x\} \leqslant e^{-x^2/2}$. Thus,

$$\begin{aligned} &\|\Delta\| \leqslant \sigma \textstyle\sum_{\ell=1}^{N} \sqrt{d_\ell} + \sigma\sqrt{2d_N}, \\ &\big\|\Psi \times_{\ell\in[N]\setminus\{k\}} g_{j\ell}^\top\big\|_2 \leqslant (1+\sqrt{2})\sigma\sqrt{d_k} \end{aligned}$$

in an event Ω_1 with at least probability $1 - \sum_{k=1}^{N} e^{-d_k} - e^{-d_N}$. Consequently, by (67), in $\Omega_1 \cap \Omega^*_{m,k-1}$,

$$\begin{aligned} &\big\|\Psi \times_{\ell=k-1}^{k-N+1} \widehat{g}_{j,\ell}^{(m)\top}\big\|_2 \big/ \big(\lambda_j \textstyle\prod_{\ell=1}^{N-1} a_{j,k-\ell}^\top \widehat{g}_{j,k-\ell}^{(m)}\big) \\ \leqslant\ & (1+\sqrt{2})\sigma\big(d_k^{1/2} + (\phi^*_{m,k-1} \wedge 1)\textstyle\sum_{\ell=1}^{N} d_\ell^{1/2}\big)\alpha_*^{1-N}/\lambda_j \\ \leqslant\ & \sqrt{8}\alpha_*^{1-N} R^{*(\mathrm{ideal})}_{jk,\phi_{m,k-1}} \end{aligned} \tag{68}$$

Substituting (68) into (66), we have, in the event $\Omega_1 \cap \Omega_{m,k-1}$,

$$\big(2 - 2|a_{jk}^\top \widehat{a}_{jk}^{(m)}|\big)^{1/2} \leqslant \psi_{m,j,k} \tag{69}$$

with

$$\psi_{m,j,k} = \max\left\{\frac{6R^{*(\mathrm{ideal})}_{jk,\phi_{m,k-1}}}{\alpha_*^{N-1}}, \frac{6\lambda_1\sqrt{r-1}}{\lambda_j\alpha_*^{N-1}} \prod_{\ell=1}^{N-1} \psi_{m,k-\ell}\right\}.$$

Consequently, $\Omega_{m,k} \subset \Omega_1 \cap \Omega^*_{m,k-1}$ and the upper bound for required number of iterations follows from the same (but much simpler) argument in Steps 2 and 3 of the proof of Theorem 3.

As for the estimation of λ_j, similar to (66), we can obtain

$$\begin{aligned} &\big|\widehat{\lambda}_j^{(m)} - \lambda_j\big| \\ \leqslant &\big\|\Psi \times_{\ell\in[N]} \widehat{b}_{j\ell}^{(m)\top}\big\|_2 + (r-1)\lambda_1 \prod_{\ell=1}^{N} \phi_{m,\ell} + \sum_{\ell=1}^{N} \phi_{m,\ell}. \end{aligned} \tag{70}$$

Then, employing similar procedures as above, we can prove the bound (34). □

References

[1] O. Alter and G. H. Golub, "Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 49, pp. 17559–17564, 2005.

[2] L. Omberg, G. H. Golub, and O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of dna microarray data from different studies," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 47, pp. 18371–18376, 2007.

[3] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[4] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. Amer. Stat. Assoc.*, vol. 108, no. 502, pp. 540–552, 2013.

[5] W. W. Sun and L. Li, "STORE: Sparse tensor response regression and neuroimaging analysis," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 4908–4944, 2017.

[6] X. Bi, A. Qu, and X. Shen, "Multilayer tensor factorization with applications to recommender systems," *Ann. Statist.*, vol. 46, no. 6B, pp. 3308–3333, Dec. 2018.

[7] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.

[8] P. D. Hoff, "Multilinear tensor regression for longitudinal relational data," *Ann. Appl. Statist.*, vol. 9, no. 3, p. 1169, Sep. 2015.

[9] E. Y. Chen and J. Fan, "Statistical inference for high-dimensional matrix-variate factor models," *J. Amer. Stat. Assoc.*, vol. 2021, pp. 1–18, Oct. 2021.

[10] R. Chen, D. Yang, and C.-H. Zhang, "Factor models for high-dimensional tensor time series," *J. Amer. Stat. Assoc.*, vol. 117, no. 537, pp. 94–116, Jan. 2022.

[11] Y. Han, R. Chen, and C.-H. Zhang, "Rank determination in tensor factor model," *Electron. J. Statist.*, vol. 16, no. 1, pp. 1726–1803, 2022.

[12] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade, "A tensor approach to learning mixed membership community models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2239–2312, Jan. 2014.

[13] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *J. Mach. Learn. Res.*, vol. 15, pp. 2773–2832, Aug. 2014.

[14] A. T. Chaganty and P. Liang, "Estimating latent-variable graphical models using moments and likelihoods," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1872–1880.

[15] B. Hao, A. Zhang, and G. Cheng, "Sparse and low-rank tensor estimation via cubic sketchings," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5927–5964, Sep. 2020.

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

[17] M. Udell and A. Townsend, "Why are big data matrices approximately low rank?" *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 144–160, Jan. 2019.

[18] J. Håstad, "Tensor rank is NP-complete," *J. Algorithms*, vol. 11, no. 4, pp. 644–654, Dec. 1990.

[19] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, pp. 1–39, Nov. 2013.

[20] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *J. Amer. Statist. Assoc.*, vol. 104, no. 486, pp. 682–693, Jan. 2009.

[21] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[22] D. Zhang and Z.-H. Zhou, "(2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition," *Neurocomputing*, vol. 69, nos. 1–3, pp. 224–231, 2005.

[23] H. Kong, L. Wang, E. K. Teoh, X. Li, J. G. Wang, and R. Venkateswarlu, "Generalized 2D principal component analysis for face image representation and recognition," *Neural Netw.*, vol. 18, nos. 5–6, pp. 585–594, 2005.

[24] Y. Pang, D. Tao, Y. Yuan, and X. Li, "Binary two-dimensional PCA," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 38, no. 4, pp. 1176–1180, Aug. 2008.

[25] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[26] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2DPCA," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 40, no. 4, pp. 1170–1175, Aug. 2010.

[27] D. Meng, Q. Zhao, and Z. Xu, "Improve robustness of sparse PCA by L_1-norm maximization," *Pattern Recognit.*, vol. 45, no. 1, pp. 487–497, Jan. 2012.

[28] R. Wang, F. Nie, X. Yang, F. Gao, and M. Yao, "Robust 2DPCA with non-greedy ℓ_1-norm maximization for image analysis," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1108–1112, May 2015.

[29] A. Anandkumar, R. Ge, and M. Janzamin, "Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates," 2014, *arXiv:1402.5180*.

[30] P. D. Hoff, "Separable covariance arrays via the tucker product, with applications to multivariate relational data," *Bayesian Anal.*, vol. 6, no. 2, pp. 179–196, Jun. 2011.

[31] B. K. Fosdick and P. D. Hoff, "Separable factor analysis with applications to mortality data," *Ann. Appl. Statist.*, vol. 8, no. 1, p. 120, Mar. 2014.

[32] E. Y. Chen, D. Xia, C. Cai, and J. Fan, "Semiparametric tensor factor analysis by iteratively projected SVD," 2020, *arXiv:2007.02404*.

[33] L. Yu, Y. He, X. Kong, and X. Zhang, "Projected estimation for large-dimensional matrix factor models," *J. Econometrics*, vol. 229, no. 1, pp. 201–217, Jul. 2022.

[34] C. Lam, "Rank determination for time series tensor factor model using correlation thresholding," LSE, London, U.K., Working Paper, 2021.

[35] E. Y. Chen, R. S. Tsay, and R. Chen, "Constrained factor models for high-dimensional matrix-variate time series," *J. Amer. Stat. Assoc.*, vol. 115, no. 530, pp. 775–793, Apr. 2020.

[36] Y. Han, R. Chen, D. Yang, and C.-H. Zhang, "Tensor factor model estimation by iterative projection," 2020, *arXiv:2006.02611*.

[37] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(R_1, R_2,..., R_N) approximation of higher-order tensors," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[38] Y. Liu, F. Shang, W. Fan, J. Cheng, and H. Cheng, "Generalized higher-order orthogonal iteration for tensor decomposition and completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1763–1771.

[39] A. Zhang and D. Xia, "Tensor SVD: Statistical and computational limits," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7311–7338, Nov. 2018.

[40] P. Comon, X. Luciani, and A. L. F. de Almeida, "Tensor decompositions, alternating least squares and other tales," *J. Chemometrics*, vol. 23, pp. 393–405, Jul. 2009.

[41] E. Richard and A. Montanari, "A statistical model for tensor PCA," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[42] P.-A. Wang and C.-J. Lu, "Tensor decomposition via simultaneous power iteration," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3665–3673.

[43] M. Wang and Y. Song, "Tensor decompositions via two-mode higher-order SVD (HOSVD)," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54. PMLR, 2017, pp. 614–622. [Online]. Available: http://proceedings.mlr.press/v54/wang17a.html

[44] Q. Le, A. Karpenko, J. Ngiam, and A. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1017–1025.

[45] A. Souloumiac, "Joint diagonalization: Is non-orthogonal always preferable to orthogonal?" in *Proc. 3rd IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 2009, pp. 305–308.

[46] F. Huang, U. N. Niranjan, M. U. Hakeem, and A. Anandkumar, "Online tensor methods for learning latent variable models," 2013, *arXiv:1309.0787*.

[47] W. W. Sun, J. Lu, H. Liu, and G. Cheng, "Provable sparse tensor decomposition," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 79, no. 3, pp. 899–916, Jun. 2017.

[48] V. Sharan and G. Valiant, "Orthogonalized ALS: A theoretically principled tensor decomposition algorithm for practical use," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3095–3104.

[49] V. Kuleshov, A. Chaganty, and P. Liang, "Tensor factorization via matrix factorization," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, vol. 38. PMLR, 2015, pp. 507–516. [Online]. Available: https://proceedings.mlr.press/v38/kuleshov15.html

[50] N. Colombo and N. Vlassis, "Tensor decomposition via joint matrix Schur decomposition," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2820–2828.

[51] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra Appl.*, vol. 18, no. 2, pp. 95–138, 1975.

[52] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul, "Minimax bounds for sparse PCA with noisy high-dimensional data," *Ann. Statist.*, vol. 41, no. 3, p. 1055, Jun. 2013.

[53] M. Brennan and G. Bresler, "Reducibility and statistical-computational gaps from secret leakage," in *Proc. Conf. Learn. Theory*, 2020, pp. 648–847.

[54] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," *Oper. Res.*, vol. 70, no. 2, pp. 1219–1237, Mar. 2022.

[55] Y. Luo and A. R. Zhang, "Low-rank tensor estimation via Riemannian Gauss–Newton: Statistical optimality and second-order convergence," 2021, *arXiv:2104.12031*.

[56] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen, "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," *Ann. Statist.*, vol. 49, no. 2, pp. 944–967, Apr. 2021.

[57] D. Xia and M. Yuan, "On polynomial time methods for exact low-rank tensor completion," *Found. Comput. Math.*, vol. 19, no. 6, pp. 1265–1313, Dec. 2019.

[58] A. Zhang and R. Han, "Optimal sparse singular value decomposition for high-dimensional high-order data," *J. Amer. Stat. Assoc.*, vol. 114, no. 528, pp. 1708–1725, 2019.

[59] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi, "Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements," *J. Mach. Learn. Res.*, vol. 23, no. 163, pp. 1–77, 2022.

[60] R. Han, R. Willett, and A. R. Zhang, "An optimal statistical and computational framework for generalized tensor estimation," *Ann. Statist.*, vol. 50, no. 1, pp. 1–29, 2022.

[61] R. Han, P. Shi, and A. R. Zhang, "Guaranteed functional tensor singular value decomposition," 2021, *arXiv:2108.04201*.

[62] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numer. Math.*, vol. 12, no. 1, pp. 99–111, 1972.

[63] V. Koltchinskii and K. Lounici, "Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance," *Annales l'Institut Henri Poincaré, Probabilités Statistiques*, vol. 52, no. 4, pp. 1976–2013, 2016.

[64] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and Banach spaces," in *Handbook of the Geometry of Banach Spaces*, vol. 1. Amsterdam, The Netherlands: Elsevier, 2001, p. 131.

Yuefeng Han received the Ph.D. degree in statistics from the University of Chicago in 2019. He is currently an Assistant Professor with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame. His research interests include tensor data analysis, high-dimensional statistical inference, and time series analysis.

Cun-Hui Zhang received the Ph.D. degree in statistics from Columbia University in 1984. He is currently a Distinguished Professor with the Department of Statistics, Rutgers University. His research interests include high-dimensional data, empirical Bayes, functional MRI, network data, semi-parametric and nonparametric methods, survival analysis, statistical inference, and probability theory. He is a fellow of the Institute of Mathematical Statistics (IMS) and the American Statistical Association (ASA).