The Annals of Statistics
2024, Vol. 52, No. 6, 2641–2667
https://doi.org/10.1214/24-AOS2412

TENSOR FACTOR MODEL ESTIMATION BY ITERATIVE PROJECTION

BY YUEFENG HAN[1,a], RONG CHEN[2,b], DAN YANG[3,d] AND CUN-HUI ZHANG[2,c]

[1]*Department of Applied and Computational Mathematics and Statistics, University of Notre Dame,* [a]*yuefeng.han@nd.edu*
[2]*Department of Statistics, Rutgers University,* [b]*rongchen@stat.rutgers.edu,* [c]*czhang@stat.rutgers.edu*
[3]*Faculty of Business and Economics, The University of Hong Kong,* [d]*dyanghku@hku.hk*

Tensor time series, which is a time series consisting of tensorial observations, has become ubiquitous. It typically exhibits high dimensionality. One approach for dimension reduction is to use a factor model structure, in a form similar to Tucker tensor decomposition, except that the time dimension is treated as a dynamic process with a time dependent structure. In this paper, we introduce two approaches to estimate such a tensor factor model by using iterative orthogonal projections of the original tensor time series. These approaches extend the existing estimation procedures and improve the estimation accuracy and convergence rate significantly as proven in our theoretical investigation. Our algorithms are similar to the higher-order orthogonal projection method for tensor decomposition, but with significant differences due to the need to unfold tensors in the iterations and the use of autocorrelation. Consequently, our analysis is significantly different from the existing ones. Computational and statistical lower bounds are derived to prove the optimality of the sample size requirement and convergence rate for the proposed methods. Simulation study is conducted to further illustrate the statistical properties of these estimators.

1. Introduction. Motivated by a diverse range of modern scientific applications, analysis of tensors or multidimensional arrays, has emerged as one of the most important and active research areas in statistics, computer science and machine learning. Large tensors are encountered in genomics (Alter and Golub (2005), Omberg, Golub and Alter (2007)), neuroimaging analysis (Zhou, Li and Zhu (2013), Sun and Li (2017)), recommender systems (Bi, Qu and Shen (2018)), computer vision (Liu et al. (2012)), community detection (Anandkumar et al. (2014)), among others. High-order tensors often bring about high dimensionality and impose significant computational challenges. For example, functional MRI produces a time series of 3-dimensional brain images, typically consisting of hundreds of thousands of voxels observed over time. Previous work has developed various tensor-based methods for independent and identically distributed (i.i.d.) tensor data or tensor data with i.i.d. noise. However, the statistical framework for general tensor time-series data is much less studied in the literature.

Factor analysis is one of the most useful tools for understanding common dependence among multidimensional outputs. Over the past decades, vector factor models have been extensively studied in the statistics and economics communities. For instance, Chamberlain and Rothschild (1983), Bai and Ng (2002), Stock and Watson (2002) and Bai (2003) developed the static factor model using principal component analysis (PCA). They assumed that the common factors must have impact on most of the time series, and weak serial dependence is allowed for the idiosyncratic noise process. Fan, Liao and Mincheva (2011, 2013), Fan, Liu and Wang (2018) established large covariance matrix estimation based on the static factor model. The static factor model has been further extended to the dynamic factor model

Received July 2021; revised December 2023.
MSC2020 subject classifications. Primary 62H25, 62H12; secondary 62R07.
Key words and phrases. High-dimensional tensor data, factor model, orthogonal projection, time series, Tucker decomposition.

in Forni et al. (2000). In the dynamic factor model, the latent factors are assumed to follow a time-series process, which is commonly taken to be a vector autoregressive process. Fan, Liao and Wang (2016) studied semiparametric factor models through projected principal component analysis. Peña and Box (1987), Pan and Yao (2008), Lam, Yao and Bathia (2011) and Lam and Yao (2012) adopted another type of factor model. They assumed that the latent factors capture all dynamics of the observed process, and thus the idiosyncratic noise process has no serial dependence. We will adopt this approach. We note that the factor process may have complex dynamic behavior, resulting in complex dynamics of the observed tensor, even with white additive noise process. Of course, when all the dynamics of the observed tensor process are "forced" to be included in the signal process induced by the factor process, situations may arise in which some factors are "weak" (or have impact on a small portion of the observed series in the tensor). This leads us to consider the "signal strength" in our investigation.

Although there have been significant efforts in developing methodologies and theories for vector factor models, there is a paucity of literature on matrix- or tensor-valued time series. Wang, Liu and Chen (2019) proposed a matrix factor model for matrix-valued time series, which explores the matrix structure. Chen, Tsay and Chen (2020) established a general framework for incorporating domain and prior knowledge in the matrix factor model through linear constraints. Chen and Chen (2022) applied the matrix factor model to the dynamic transport network. Chen and Fan (2023) developed an inferential theory of the matrix factor model under a different setting from that in Wang, Liu and Chen (2019). Chang et al. (2023), Han and Zhang (2023), Han, Zhang and Chen (2021) studied factor models with CP type low rank structures.

Recently, Chen, Yang and Zhang (2022a) introduced a factor approach for analyzing high-dimensional dynamic tensor time series in the form

$$\mathcal{X}_t = \mathcal{M}_t + \mathcal{E}_t, \tag{1.1}$$

where $\mathcal{X}_1, \ldots, \mathcal{X}_T \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ are the observed tensor time series, $\mathcal{M}_t$ and $\mathcal{E}_t$ are the corresponding signal and noise components of $\mathcal{X}_t$, respectively. The goal is to estimate the unknown signal tensor $\mathcal{M}_t$ from the tensor time series data. Following Lam and Yao (2012), it is assumed that the signal tensor accommodates all dynamics, making the idiosyncratic noise $\mathcal{E}_t$ uncorrelated (white) across time. It is further assumed that $\mathcal{M}_t$ lives in a lower-dimensional space and has certain multilinear decomposition. Specifically, we assume that $\mathcal{M}_t$ satisfies a Tucker-type decomposition and model (1.1) can be written as

$$\mathcal{X}_t = \mathcal{F}_t \times_1 A_1 \times_2 \ldots \times_K A_K + \mathcal{E}_t, \tag{1.2}$$

where A_k is the deterministic loading matrix of size $d_k \times r_k$ and $r_k \ll d_k$, and the core tensor $\mathcal{F}_t$ itself is a latent tensor factor process of dimension $r_1 \times \cdots \times r_K$. Here, the k-mode product of $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ with a matrix $U \in \mathbb{R}^{d_k' \times d_k}$, denoted as $\mathcal{X} \times_k U$, is an order K-tensor of size $d_1 \times \cdots \times d_{k-1} \times d_k' \times d_{k+1} \times \cdots \times d_K$ such that

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

The core tensor $\mathcal{F}_t$ is usually much smaller than $\mathcal{X}_t$ in dimension. This structure provides an effective dimension reduction, as all the comovements of individual time series in $\mathcal{X}_t$ are driven by $\mathcal{F}_t$. Without loss of generality, assume that A_k is of rank $r_k \ll d_k$. It should be noted that vector and matrix factor models can be viewed as special cases of our model since a vector time series is a tensor time series composed of a single fiber ($K = 1$), and a matrix times series is one composed of a single slice ($K = 2$).

Chen, Yang and Zhang (2022a) proposed two estimation procedures, namely TOPUP and TIPUP, for estimating the column space spanned by the loading matrix A_k, for $k = 1, \ldots, K$. The two procedures are based on different autocross-product operations of the observed tensors $\mathcal{X}_t$ to accumulate information, but they both utilize the assumption that the noise $\mathcal{E}_t$ and $\mathcal{E}_{t-h}, h > 0$ are uncorrelated. The convergence rates of their estimators critically depend on $d = d_1 d_2 \ldots d_K$, a potentially very large number as $d_k, k = 1, \ldots, K$, are large. Often a large T, the length of the time series, is required for accurate estimation of the loading spaces.

In this paper, we propose extensions of the TOPUP and TIPUP procedures, motivated by the following observation. Suppose that the loading matrices A_k are orthonormal with $A_k^\top A_k = I$, and we are given $A_2, \ldots, A_K$. Let

$$\mathcal{Z}_t = \mathcal{X}_t \times_2 A_2^\top \times_3 \ldots \times_K A_K^\top; \quad \text{and} \quad \mathcal{E}_t^* = \mathcal{E}_t \times_2 A_2^\top \times_3 \ldots \times_K A_K^\top.$$

Then (1.2) leads to

$$\mathcal{Z}_t = \mathcal{F}_t \times_1 A_1 + \mathcal{E}_t^*, \tag{1.3}$$

where $\mathcal{Z}_t$ is a $d_1 \times r_2 \times \cdots \times r_K$ tensor. Since $r_k \ll d_k$, $\mathcal{Z}_t$ is a much smaller tensor than $\mathcal{X}_t$. Under proper conditions on the combined noise tensor $\mathcal{E}_t^*$, the estimation of the loading space of A_1 based on $\mathcal{Z}_t$ can be made significantly more accurate, as the convergence rate now depends on $d_1 r_2 \ldots r_K$ rather than $d_1 d_2 \ldots d_K$.

Of course, in practice we do not know $A_2, \ldots, A_K$. Similar to backfitting algorithms, we propose an iterative algorithm. With a proper initial value, we iteratively estimate the loading space of A_k at iteration j based on

$$\mathcal{Z}_{t,k}^{(j)} = \mathcal{X}_t \times_1 \widehat{A}_1^{(j)\top} \times_2 \ldots \times_{k-1} \widehat{A}_{k-1}^{(j)\top} \times_{k+1} \widehat{A}_{k+1}^{(j-1)\top} \times_{k+2} \ldots \times_K \widehat{A}_K^{(j-1)\top},$$

using the estimate $\widehat{A}_{k'}^{(j-1)}, k < k' \le K$ obtained in the previous iteration and the estimate $\widehat{A}_{k'}^{(j)}, 1 \le k' < k$, obtained in the current iteration. Our theoretical investigation shows that the iterative procedures for estimating A_1 can achieve the convergence rate as if all $A_2, \ldots, A_K$ are known and we indeed observe $\mathcal{Z}_t$ that follows model (1.3). We call the procedure iTOPUP and iTIPUP, based on the matrix unfolding mechanism used, corresponding to TOPUP and TIPUP procedures. To be more specific, our algorithms have two steps: (i) We first use the estimated column space of factor loading matrices of TOPUP (resp., TIPUP) to construct the initial estimate of factor loading spaces; (ii) We then iteratively perform matrix unfolding of the autocross-moments of much smaller tensors $\mathcal{Z}_{t,k}^{(j)}$ to obtain the final estimator.

We note that the iterative procedure is related to higher order orthogonal iteration (HOOI) that has been widely studied in the literature; see, for example, De Lathauwer, De Moor and Vandewalle (2000), Sheehan and Saad (2007), Liu et al. (2014), Zhang and Xia (2018), among others. However, most of the existing works are not designed for tensor time series. They do not consider the special role of the time mode nor the covariance structure in the time direction. Typically, HOOI treats the signal part as fixed or deterministic. In this paper, we treat the signal as dynamic in the sense that the core tensor $\mathcal{F}_t$ in (1.2) is dynamic and the relationship between $\mathcal{F}_t$ and the lagged $\mathcal{F}_{t-h}$ is of interest. Our setting requires special treatment although each iteration of our iterative procedures also consists of power up and orthogonal projection operations. While HOOI applies the SVD directly to the matrix unfolding of the iteratively projected data, in our approach the SVD is applied to the matrix unfolding of the outer and inner autocross-product of the iteratively projected data, respectively in iTOPUP and iTIPUP. Although the iTOPUP algorithm proposed here can be reformulated as a twist of HOOI on the autocross-moment tensor, the iTIPUP algorithm is different and cannot be recast equivalently as HOOI. More importantly, the theoretical analysis and theoretical properties of the estimators are fundamentally different from those of HOOI, due to the dynamic

structure of tensor time series and the need to use the autocross-product operation between the SVD and data projection in each iteration. Different concentration inequalities are derived to study the performance bounds.

In this paper, we establish upper bounds on the estimation errors for both the iTOPUP and the iTIPUP, which are much sharper than the respective theoretical guarantees for TOPUP and TIPUP, demonstrating the benefits of using iterative projection. It is also shown that the number of iterations needed for convergence is of order no greater than $\log(d)$. We mainly focus on the cases where the tensor dimensions are large and of similar order. We also cover the cases where the ranks of the tensor factor process increase with the dimensions of the tensor time series.

Chen, Yang and Zhang (2022a) showed that the TIPUP has a faster convergence rate in estimation error than the TOPUP, under a mild condition on the level of signal cancellation. In contrast, the theoretically guaranteed rate of convergence for the iTOPUP in this paper is of the same order or even faster than that for the iTIPUP under certain regularity conditions. Our results also suggest an interesting phenomenon. Using the iterative procedures, we find that the increase in either dimension or sample size can improve the estimation of the factor loading space of the tensor factor model with the tensor order $K \geq 2$. We believe that such a super convergence rate is new in the literature. Specifically, under proper regularity conditions, the convergence rate of the iterative procedures for estimating the space of A_k is $O_{\mathbb{P}}(T^{-1/2}d_{-k}^{-1/2})$, where $d_{-k} = \prod_{j \neq k} d_j$, while the existing rate for noniterative procedures is $O_{\mathbb{P}}(T^{-1/2})$ for the vector factor model (Lam, Yao and Bathia (2011)) and the matrix/tensor factor models (Wang, Liu and Chen (2019), Chen, Yang and Zhang (2022a)). While the increase in the dimensions d_k $(k = 1, \ldots, K)$ does not improve the performance of the noniterative estimators, it significantly improves that of the proposed iterative estimators.

In addition, we establish the computational lower bound for the estimation of the loading spaces of tensor factor models under the hardness assumption of certain instances of hypergraphic planted clique detection problem. It shows that the sample size requirement (or signal to noise ratio condition) needed for using the TIPUP estimate as the initial values for the iterative procedures is unavoidable for any computationally manageable estimation procedure to achieve consistency, although the iterative procedures have faster convergence rates. Furthermore, we provide a statistical lower bound that matches the convergence rates of our iterative procedures under certain conditions, revealing a different effect of the ranks r_k $(k = 1, \ldots, K)$ compared to tensor Tucker decomposition (Zhang and Xia (2018)).

Related work. We close this section by highlighting several recent papers on related topics. First, we draw attention to the work of Foster (1996), Fan, Liao and Wang (2016) and Chen et al. (2024). Chen et al. (2024) adopts an estimation precedure composed of a spectral initialization followed by an iterative refinement step, so that our methods are related to theirs. However, due to the differences in problem setting and model assumptions, their estimation procedures, performance bounds and analytic techniques are all significantly different from ours. Foster (1996), Fan, Liao and Wang (2016) use the projection to the space spanned by the sieve bases without iteration. Rogers, Li and Russell (2013) assumes the tensor factor model in (1.2), with an additional specific AR structure on the dynamic of the factor process. The additional model structure in their paper led to an EM type of estimation approach, quite different from the approach we develop here. Wang, Zheng and Li (2024) concerns low rank tensor AR model and uses a nuclear norm penalty to enforce the low rank structure and optimization algorithms for estimation, again quite different from our approach.

The paper is organized as follows. Section 2.1 introduces basic notation and preliminaries of tensor analysis. We present the tensor factor model and the iTOPUP and iTIPUP procedures in Sections 2.2 and 2.3. Theoretical properties of the iTOPUP and iTIPUP are investigated in Section 3. Section 5 provides a brief summary. Numerical comparison of our iterative

procedures and other methods, and all technical details are relegated to the Supplementary Material (Han et al. (2024)).

2. Tensor factor model by orthogonal iteration.

2.1. *Notation and preliminaries for tensor analysis*. Throughout this paper, for a vector $x = (x_1, \ldots, x_p)^\top$, define $\|x\|_q = (x_1^q + \cdots + x_p^q)^{1/q}$, $q \geq 1$. For a matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, write the SVD as $A = U\Sigma V^\top$, where $\Sigma = \operatorname{diag}(\sigma_1(A), \sigma_2(A), \ldots, \sigma_{\min\{m,n\}}(A))$, with singular values $\sigma_{\max}(A) = \sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A) \geq 0$ in descending order. The matrix spectral norm is denoted as $\|A\|_{\mathrm{S}} = \sigma_1(A)$. Write $\sigma_{\min}(A)$ the smallest nontrivial singular value of A. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ (resp., $a_n \asymp b_n$) if there exists a constant C such that $|a_n| \leq C|b_n|$ (resp., $1/C \leq a_n/b_n \leq C$) for all sufficiently large n, and write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. Write $a_n \lesssim b_n$ (resp., $a_n \gtrsim b_n$) if there exist a constant C such that $a_n \leq Cb_n$ (resp., $a_n \geq Cb_n$). Write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. We use $C, C_1, c, c_1, \ldots$ to denote generic constants, whose actual values may vary from line to line.

For any two $m \times r$ matrices with orthonormal columns, say, U and $\widehat{U}$, suppose the singular values of $U^\top\widehat{U}$ are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. A natural measure of distance between the column spaces of U and $\widehat{U}$ is then

$$\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathrm{S}} = \sqrt{1-\sigma_r^2}, \tag{2.1}$$

which equals to the sine of the largest principle angle between the column spaces of U and $\widehat{U}$. For any two matrices $A \in \mathbb{R}^{m_1\times r_1}$, $B \in \mathbb{R}^{m_2\times r_2}$, denote the Kronecker product $\odot$ as $A \odot B \in \mathbb{R}^{m_1m_2\times r_1r_2}$. 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

$$(\mathcal{A}\otimes\mathcal{B})_{i_1,\ldots,i_K,j_1,\ldots,j_N} = (\mathcal{A})_{i_1,\ldots,i_K}(\mathcal{B})_{j_1,\ldots,j_N}.$$

Let $\operatorname{vec}(\cdot)$ be the vectorization of matrices and tensors. The mode-k unfolding (or matricization) is defined as $\operatorname{mat}_k(\mathcal{A})$, which maps a tensor $\mathcal{A}$ to a matrix $\operatorname{mat}_k(\mathcal{A}) \in \mathbb{R}^{m_k\times m_{-k}}$ where $m_{-k} = \prod_{j\neq k}^K m_j$. For example, if $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times m_3}$, then

$$\big(\operatorname{mat}_1(\mathcal{A})\big)_{i,(j+m_2(k-1))} = \big(\operatorname{mat}_2(\mathcal{A})\big)_{j,(k+m_3(i-1))} = \big(\operatorname{mat}_3(\mathcal{A})\big)_{k,(i+m_1(j-1))} = \mathcal{A}_{ijk}.$$

For tensor $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$, the Hilbert–Schmidt norm is defined as

$$\|\mathcal{A}\|_{\mathrm{HS}} = \sqrt{\sum_{i_1=1}^{m_1}\cdots\sum_{i_K=1}^{m_K}(\mathcal{A})_{i_1,\ldots,i_K}^2}.$$

For a matrix, the Hilbert–Schmidt norm is just the Frobenius norm. Define the tensor operator norm for an order-4 tensor $\mathcal{A} \in \mathbb{R}^{m_1\times m_2\times m_3\times m_4}$,

$$\|\mathcal{A}\|_{\mathrm{op}} = \max\Big\{\sum_{i_1,i_2,i_3,i_4} u_{i_1,i_2}\cdot u_{i_3,i_4}\cdot(\mathcal{A})_{i_1,i_2,i_3,i_4} : \|U_1\|_{\mathrm{HS}} = \|U_2\|_{\mathrm{HS}} = 1\Big\},$$

where $U_1 = (u_{i_1,i_2}) \in \mathbb{R}^{m_1\times m_2}$ and $U_2 = (u_{i_3,i_4}) \in \mathbb{R}^{m_3\times m_4}$.

2.2. *Tensor factor model*. Again, we consider as in (1.2),

$$\mathcal{X}_t = \mathcal{F}_t \times_1 A_1 \times_2 \ldots \times_K A_K + \mathcal{E}_t.$$

Without loss of generality, assume that A_k is of rank r_k. A_k is not necessarily orthonormal, which is different from the classical Tucker decomposition (Tucker (1966)). Model (1.2) is

unchanged if we replace $(A_1, \ldots, A_K, \mathcal{F}_t)$ by $(A_1H_1, \ldots, A_KH_K, \mathcal{F}_t \times_{k=1}^K H_k^{-1})$ for any invertible $r_k \times r_k$ matrix H_k. Although $(A_1, \ldots, A_K, \mathcal{F}_t)$ are not uniquely determined, the factor loading space, that is, the linear space spanned by the columns of A_k, is uniquely defined. Denote the orthogonal projection to the column space of A_k as

$$P_k = P_{A_k} = A_k(A_k^\top A_k)^{-1}A_k^\top = U_kU_k^\top, \tag{2.2}$$

where U_k is the left singular matrix in the SVD $A_k = U_k\Lambda_kV_k^\top$. We use P_k to represent the factor loading space of A_k. Thus, our objective is to estimate P_k.

The canonical representation of the tensor times series (1.2) is written as

$$\mathcal{X}_t = \mathcal{F}_t^{(\text{cano})} \times_{k=1}^K U_k + \mathcal{E}_t,$$

where the diagonal and right singular matrices of A_k are absorbed into the canonical core tensor $\mathcal{F}_t^{(\text{cano})} = \mathcal{F}_t \times_{k=1}^K (\Lambda_kV_k^\top)$. In this canonical form, the loading matrices U_k are identifiable up to a rotation in general and up to a permutation and sign changes of the columns of U_k when the singular values are all distinct in the population version of the TOPUP or TIPUP methods, as we describe in Section 2.3 below. In what follows, we may identify the tensor time series in its canonical form, that is, $A_k = U_k$, without explicit declaration.

We do not impose any specific structure for the dynamics of the core tensor factor process $\mathcal{F}_t \in \mathbb{R}^{r_1\times\cdots\times r_K}$ beyond the independence between the core process and the noise process, and we do not require any additional structure on the correlation among different time-series fibers of the noise process $\mathcal{E}_t$. Because of this generality, our estimator is based on the tensor version of the lagged sample cross-product $\widehat{\Sigma}_h$, $h = 1, \ldots, h_0$, where

$$\widehat{\Sigma}_h = \widehat{\Sigma}_h(\mathcal{X}_{1:T}) = \sum_{t=h+1}^T \frac{\mathcal{X}_{t-h} \otimes \mathcal{X}_t}{T-h} \in \mathbb{R}^{d_1\times\cdots\times d_K\times d_1\times\cdots\times d_K} \tag{2.3}$$

is an order-$2K$ tensor. The population version of this tensor autocovariance is

$$\Sigma_h = \mathbb{E}\left(\sum_{t=h+1}^T \frac{\mathcal{X}_{t-h} \otimes \mathcal{X}_t}{T-h}\right) = \mathbb{E}\left(\sum_{t=h+1}^T \frac{\mathcal{M}_{t-h} \otimes \mathcal{M}_t}{T-h}\right).$$

Because $\mathcal{M}_t = \mathcal{M}_t \times_{k=1}^K P_k$ for all t,

$$\Sigma_h = \Sigma_h \times_{k=1}^{2K} P_k = \mathbb{E}\left(\sum_{t=h+1}^T \frac{\mathcal{F}_{t-h} \otimes \mathcal{F}_t}{T-h}\right) \times_{k=1}^{2K} P_kA_k,$$

with the notation $A_k = A_{k-K}$ and $P_k = P_{k-K}$ for all $k > K$.

2.3. *Estimating procedures.* In this paper, we consider iterative estimation procedures to achieve sharper convergence rates than the TOPUP and TIPUP procedures proposed in Chen, Yang and Zhang (2022a). We start with a quick description of their procedures as they serve as the starting point of our proposed iTOPUP and iTIPUP procedures. Note that the procedure in Chen and Chen (2022) and Wang, Liu and Chen (2019) is the noniterative TOPUP.

(i) *Time-series Outer Product Unfolding Procedure (TOPUP)*

Let $\widehat{\Sigma}_h$ be the sample autocovariance of the data $\mathcal{X}_{1:T} = (\mathcal{X}_1, \ldots, \mathcal{X}_T)$ as in (2.3). Define

$$\text{TOPUP}_k = \big(\text{mat}_k(\widehat{\Sigma}_h), h = 1, \ldots, h_0\big), \tag{2.4}$$

as a $d_k \times (dd_{-k}h_0)$ matrix, where $d = \prod_{k=1}^K d_k$, $d_{-k} = d/d_k$ and h_0 is a predetermined positive integer. Here, we note that TOPUP_k is a function mapping a tensor time series to a matrix. In TOPUP_k, the information from different time lags is accumulated, which is useful

especially when the sample size T is small. A relatively small h_0 is typically used, since the autocorrelation is often at its strongest with small time lags; see Remark 3.10.

The TOPUP method performs SVD of (2.4) to obtain the truncated left singular matrices

$$\widehat{U}_k\text{-TOPUP}(\mathcal{X}_{1:T}, m) = \text{LSVD}_m\big(\text{mat}_k\big(\widehat{\Sigma}_h(\mathcal{X}_{1:T})\big), h = 1, \ldots, h_0\big), \tag{2.5}$$

where LSVD_m stands for the left singular matrix composed of the first m left singular vectors corresponding to the largest m singular values. Here, $\widehat{U}_k$-TOPUP is treated as an operator that maps a noisy tensor time series to a matrix of m columns as an estimate of the mode-k singular space of the low-rank signal tensor time series.

By (1.2) and (2.3), the expectation of (2.4) satisfies

$$\begin{aligned}&\mathbb{E}[\text{TOPUP}_k]\\ &= A_k \,\text{mat}_k\Bigg(\sum_{t=h+1}^{T} \mathbb{E}\bigg(\frac{\mathcal{F}_{t-h} \otimes \mathcal{F}_t}{T-h}\bigg) \times_{l=1}^{k-1} A_l \times_{l=k+1}^{2K} A_l, h = 1, \ldots, h_0\Bigg),\end{aligned} \tag{2.6}$$

so that the TOPUP is expected to be consistent in estimating the column space of A_k.

(ii) *Time-series Inner Product Unfolding Procedure (TIPUP)*

Similar to (2.4), define a $d_k \times (d_k h_0)$ matrix as

$$\text{TIPUP}_k = \Bigg(\sum_{t=h+1}^{T} \frac{\text{mat}_k(\mathcal{X}_{t-h})\text{mat}_k^\top(\mathcal{X}_t)}{T-h}, h = 1, \ldots, h_0\Bigg), \tag{2.7}$$

which replaces the tensor product by the inner product through (2.3) in (2.4). The TIPUP method performs SVD:

$$\widehat{U}_k\text{-TIPUP}(\mathcal{X}_{1:T}, m) = \text{LSVD}_m\Bigg(\sum_{t=h+1}^{T} \frac{\text{mat}_k(\mathcal{X}_{t-h})\text{mat}_k^\top(\mathcal{X}_t)}{T-h}, h = 1, \ldots, h_0\Bigg), \tag{2.8}$$

for $k = 1, \ldots, K$. Again, $\widehat{U}_k$-TIPUP is treated as an operator. We note that the TOPUP method in (2.5) utilizes the entire autocross-product tensor by applying the SVD to its mode k unfolding, whereas the TIPUP only utilizes a matrix-valued linear mapping of the autocross-product tensor by first taking the model-k' trace operation for all $k' \neq k$. The trace operation cancels the noise but also possibly some signal.

(iii) *iTOPUP and iTIPUP*

Next, we describe a generic iterative procedure under the motivation described in Section 1. Its pseudocode is provided in Algorithm 1. It incorporates two estimators/operators $\widehat{U}_k$-INIT and $\widehat{U}_k$-ITER that map a tensor time series to an estimate of the loading matrix U_k. Respectively, they stand for the procedures used for initialization and iteration. The $\widehat{U}_k$-TOPUP and $\widehat{U}_k$-TIPUP operators in (2.5) and (2.8) are examples of such operators.

When we use the $\widehat{U}_k$-TOPUP operator (2.5) for both $\widehat{U}_k$-INIT and $\widehat{U}_k$-ITER in Algorithm 1, it will be called iTOPUP procedure. Similarly, iTIPUP uses $\widehat{U}_k$-TIPUP operator (2.8) for both $\widehat{U}_k$-INIT and $\widehat{U}_k$-ITER. Besides these two versions, we may also use $\widehat{U}_k$-TIPUP for $\widehat{U}_k$-INIT and $\widehat{U}_k$-TOPUP for $\widehat{U}_k$-ITER, named as TIPUP-iTOPUP. Similarly, TOPUP-iTIPUP uses $\widehat{U}_k$-TOPUP as $\widehat{U}_k$-INIT and $\widehat{U}_k$-TIPUP as $\widehat{U}_k$-ITER. These variants are sometimes useful, because TOPUP and TIPUP have different theoretical properties as the initializer or for iteration, as we will discuss in Section 3. Other estimators of the loading spaces based on the tensor time series can also be used in place of $\widehat{U}_k$-INIT and $\widehat{U}_k$-ITER, such as the conventional high order SVD for tensor decomposition, which we refer to as the Unfolding Procedure (UP), that simply performs SVD of the matricization along the appropriate mode of the $K+1$ order tensor $(\mathcal{X}_1, \ldots, \mathcal{X}_T)$ with time dimension as the additional $(K+1)$th mode.

Algorithm 1 A generic iterative algorithm

1: Input: $\mathcal{X}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ for $t = 1, \ldots, T$, r_k for all $k = 1, .., K$, the tolerance parameter $\epsilon > 0$, the maximum number of iterations J, and the $\widehat{U}_k$-INIT and $\widehat{U}_k$-ITER operators.

2: Let $j = 0$, initiate via applying $\widehat{U}_k$-INIT on $\{\mathcal{X}_{1:T}\}$, for $k = 1, \ldots, K$, to obtain

$$\widehat{U}_k^{(0)} = \widehat{U}_k\text{-INIT}_k(\mathcal{X}_{1:T}, r_k).$$

3: **repeat**

4: Let $j = j + 1$. At the jth iteration, for $k = 1, \ldots, K$, given previous estimates $(\widehat{U}_{k+1}^{(j-1)}, \ldots, \widehat{U}_K^{(j-1)})$ and $(\widehat{U}_1^{(j)}, \ldots, \widehat{U}_{k-1}^{(j)})$, sequentially calculate

$$\mathcal{Z}_{t,k}^{(j)} = \mathcal{X}_t \times_1 \big(\widehat{U}_1^{(j)}\big)^\top \times_2 \cdots \times_{k-1} \big(\widehat{U}_{k-1}^{(j)}\big)^\top \times_{k+1} \big(\widehat{U}_{k+1}^{(j-1)}\big)^\top \times_{k+2} \cdots \times_K \big(\widehat{U}_K^{(j-1)}\big)^\top,$$

for $t = 1, \ldots, T$. Perform $\widehat{U}_k$-ITER on the new tensor time series $\mathcal{Z}_{1:T,k}^{(j)} = (\mathcal{Z}_{1,k}^{(j)}, \ldots, \mathcal{Z}_{T,k}^{(j)})$.

$$\widehat{U}_k^{(j)} = \widehat{U}_k\text{-ITER}_k(\mathcal{Z}_{1:T,k}^{(j)}, r_k).$$

5: **until** $j = J$ or

$$\max_{1 \le k \le K} \|\widehat{U}_k^{(j)}(\widehat{U}_k^{(j)})^\top - \widehat{U}_k^{(j-1)}(\widehat{U}_k^{(j-1)})^\top\|_{\mathrm{S}} \le \epsilon,$$

6: Estimate and output:

$$\begin{aligned}
\widehat{U}_k^{\text{iFinal}} &= \widehat{U}_k^{(j)}, \quad k = 1, \ldots, K, \\
\widehat{P}_k^{\text{iFinal}} &= \widehat{U}_k^{\text{iFinal}}\big(\widehat{U}_k^{\text{iFinal}}\big)^\top, \quad k = 1, \ldots, K, \\
\widehat{\mathcal{F}}_t^{\text{iFinal}} &= \mathcal{X}_t \times_{k=1}^K \big(\widehat{U}_k^{\text{iFinal}}\big)^\top, \quad t = 1, \ldots, T, \\
\widehat{\mathcal{E}}_t^{\text{iFinal}} &= \mathcal{X}_t - \mathcal{X}_t \times_1 \widehat{P}_1^{\text{iFinal}} \times_2 \cdots \times_K \widehat{P}_K^{\text{iFinal}}, \quad t = 1, \ldots, T.
\end{aligned}$$

REMARK 2.1. While Algorithm 1 resembles an HOOI-type iteration of the orthogonal projection and singular matrix estimation methods, the proposed iTOPUP and iTIPUP are significantly different from HOOI, which iterates the operations of

$$\text{orthogonal projection} \to \text{matrix unfolding} \to \text{SVD}.$$

In both iTOPUP and iTIPUP, each iteration carries out the operations

$$\text{orthogonal projection} \to \text{autocovariance} \to \text{matrix unfolding} \to \text{SVD}. \tag{2.9}$$

As the outer product is taken with TOPUP_k in (2.4), its orthogonal projection and autocovariance operations are exchangeable, so that we can write

$$\text{iTOPUP} = \text{HOOI}(\hat{\Sigma}_h, h = 1, \ldots, h_0)$$

as long as the HOOI is modified by applying $U_\ell^{(j)}$ to both mode ℓ and mode $K + \ell$, $\ell \neq k$ in the projection operation and leaving alone the $(2K + 1)$th mode in the lags $1 : h_0$ throughout. However, for iTIPUP, the orthogonal projection and autocovariance operations in (2.9) are not exchangeable as the projections are sandwiched inside the autocovariance. Needless to say, the analysis of iTOPUP and iTIPUP is much more difficult than the conventional HOOI with i.i.d. assumption due to the involvement of the autocovarinace operations in the time-axis in the iterations.

REMARK 2.2 (Rank determination). Here, the estimators are constructed with given ranks $r_1, \dots, r_K$, though in theoretical analysis they are allowed to diverge. In practice, existing procedures for rank determination in the vector factor model, including the information criteria approach (Bai and Ng (2002, 2007), Hallin and Liška (2007)) and ratio of eigenvalues approach (Lam and Yao (2012), Ahn and Horenstein (2013)) can be extended to the tensor factor model by treating $d_1 \times \cdots \times d_k$ tensors as d-dimensional vectors, $d = \prod_{k=1}^K d_k$.

3. Theoretical properties. In this section, we present some theoretical properties of the iterative procedures. We first present the additional notation needed for the discussion, and then the error bounds for the iterative estimators under a minimum condition on the error process $\mathcal{E}_t$ in the model. These error bounds are quite general and cover many different models. To help decipher the general results, we present two concrete signal process models (or general sets of assumptions) with simpler and more explicit convergence rates.

3.1. *Notation.* Let $\overline{\mathbb{E}}[\cdot] = \mathbb{E}[\cdot|\{\mathcal{F}_1, \dots, \mathcal{F}_T\}]$. Define $d = \prod_{k=1}^K d_k$, $d_{-k} = d/d_k$, $r = \prod_{k=1}^K r_k$ and $r_{-k} = r/r_k$. Define order-4 tensors

$$\Theta_{k,h} = \sum_{t=h+1}^{T} \frac{\operatorname{mat}_k(\mathcal{M}_{t-h}) \otimes \operatorname{mat}_k(\mathcal{M}_t)}{T-h} \in \mathbb{R}^{d_k \times d_{-k} \times d_k \times d_{-k}},$$

$$\Phi_{k,h} = \sum_{t=h+1}^{T} \frac{\operatorname{mat}_k(\mathcal{F}_{t-h}) \otimes \operatorname{mat}_k(\mathcal{F}_t)}{T-h} \in \mathbb{R}^{r_k \times r_{-k} \times r_k \times r_{-k}}, \tag{3.1}$$

$$\Phi_{k,h}^{(\mathrm{cano})} = \sum_{t=h+1}^{T} \frac{\operatorname{mat}_k(\mathcal{M}_{t-h} \times_{k=1}^K U_k^\top) \otimes \operatorname{mat}_k(\mathcal{M}_t \times_{k=1}^K U_k^\top)}{T-h} \in \mathbb{R}^{r_k \times r_{-k} \times r_k \times r_{-k}},$$

with U_k from the SVD $A_k = U_k \Lambda_k V_k^\top$. We view $\Phi_{k,h}^{(\mathrm{cano})}$ as the canonical version of the autocovariance of the factor process. The noiseless version of the matrix TOPUP_k in (2.4) is

$$\operatorname{mat}_1(\Theta_{k,1:h_0}) = \overline{\mathbb{E}}[\mathrm{TOPUP}_k] \in \mathbb{R}^{d_k \times (dd_{-k}h_0)}, \tag{3.2}$$

with $\Theta_{k,1:h_0} = (\Theta_{k,h}, h = 1, \dots, h_0)$. The canonical factor version of (3.2) is $\operatorname{mat}_1(\Phi_{k,1:h_0}^{(\mathrm{cano})}) \in \mathbb{R}^{r_k \times (rr_{-k}h_0)}$ with $\Phi_{k,1:h_0}^{(\mathrm{cano})} = (\Phi_{k,h}^{(\mathrm{cano})}, h = 1, \dots, h_0) \in \mathbb{R}^{r_k \times r_{-k} \times r_k \times r_{-k} \times h_0}$. Similarly, define

$$\Theta_{k,h}^* = \sum_{t=h+1}^{T} \frac{\operatorname{mat}_k(\mathcal{M}_{t-h}) \operatorname{mat}_k^\top(\mathcal{M}_t)}{T-h} \in \mathbb{R}^{d_k \times d_k},$$

$$\Phi_{k,h}^* = \sum_{t=h+1}^{T} \frac{\operatorname{mat}_k(\mathcal{F}_{t-h}) \operatorname{mat}_k^\top(\mathcal{F}_t)}{T-h} \in \mathbb{R}^{r_k \times r_k}, \tag{3.3}$$

$$\begin{aligned}\Phi_{k,h}^{*(\mathrm{cano})} &= U_k^\top \Theta_{k,h}^* U_k \\ &= \sum_{t=h+1}^{T} \frac{\operatorname{mat}_k(\mathcal{M}_{t-h} \times_{k=1}^K U_k^\top) \operatorname{mat}_k^\top(\mathcal{M}_t \times_{k=1}^K U_k^\top)}{T-h} \in \mathbb{R}^{r_k \times r_k}.\end{aligned}$$

The noiseless version of (2.7) is

$$\Theta_{k,1:h_0}^* = (\Theta_{k,h}^*, h = 1, \dots, h_0) = \overline{\mathbb{E}}[\mathrm{TIPUP}_k] \in \mathbb{R}^{d_k \times (d_k h_0)}, \tag{3.4}$$

and its canonical factor version is $\Phi_{k,1:h_0}^{*(\mathrm{cano})} = (\Phi_{k,h}^{*(\mathrm{cano})}, h = 1, \dots, h_0) \in \mathbb{R}^{r_k \times (r_k h_0)}$. Let $\tau_{k,m}$ be the mth singular value of the noiseless version of the TOPUP_k matrix,

$$\tau_{k,m} = \sigma_m\big(\overline{\mathbb{E}}[\mathrm{TOPUP}_k]\big) = \sigma_m\big(\operatorname{mat}_1(\Theta_{k,1:h_0})\big) = \sigma_m\big(\operatorname{mat}_1\big(\Phi_{k,1:h_0}^{(\mathrm{cano})}\big)\big).$$

The signal strength for iTOPUP can be characterized as

$$\lambda_k = \sqrt{h_0^{-1/2}\tau_{k,r_k}}. \tag{3.5}$$

Similarly, let

$$\tau^*_{k,m} = \sigma_m\big(\overline{\mathbb{E}}(\mathrm{TIPUP}_k)\big) = \sigma_m\big(\Theta^*_{k,1:h_0}\big) = \sigma_m\big(\Phi^{*(\mathrm{cano})}_{k,1:h_0}\big).$$

The signal strength for iTIPUP can be characterized as

$$\lambda^*_k = \sqrt{h_0^{-1/2}\tau^*_{k,r_k}}. \tag{3.6}$$

We note that by (3.3) and the Cauchy–Schwarz inequality,

$$\lambda^{*2}_k \le h_0^{-1/2}\big\|\Theta^*_{k,1:h_0}\big\|_{\mathrm{S}} \le \max_{h\le h_0}\big\|\Theta^*_{k,h}\big\|_{\mathrm{S}} \le \big\|\Theta^*_{k,0}\big\|_{\mathrm{S}}/(1-h_0/T).$$

3.2. *General error bounds.* Our general error bounds for the proposed iTOPUP and iTIPUP are established under the following assumption for the error process.

ASSUMPTION 1. The error process $\mathcal{E}_t$ are independent Gaussian tensors conditionally on the factor process $\{\mathcal{F}_t, t\in\mathbb{Z}\}$. In addition, there exists some constant $\sigma > 0$, such that

$$\overline{\mathbb{E}}\big(u^\top \operatorname{vec}(\mathcal{E}_t)\big)^2 \le \sigma^2\|u\|_2^2, \quad u\in\mathbb{R}^d.$$

Assumption 1 is used in Chen, Yang and Zhang (2022a) for the theoretical investigation of the noniterative TIPUP and TOPUP, and is similar to those on the noise imposed in Lam, Yao and Bathia (2011), Lam and Yao (2012). The normality assumption, which ensures fast convergence rates in our analysis, is imposed for technical convenience. It accommodates general patterns of dependence among individual time-series fibers, but also allows a presentation of the main results with manageable analytical complexity. In fact, direct extension is visible in our analysis under the sub-Gaussian and even more general tail probability conditions. Under Assumption 1, the magnitude of the noise can be measured by the dimension d_k before the projection and by the rank r_k after the projection. The main theorems (Theorems 3.1, 3.2 and 3.3) in this section are based on this assumption on the noise alone, and cover all thereafter discussed settings of the signal $\mathcal{M}_t$.

Let us first study the behavior of iTOPUP procedure. By Chen, Yang and Zhang (2022a), the risk $\overline{\mathbb{E}}[\|\widehat{U}_k^{(0)}\widehat{U}_k^{(0)\top} - U_kU_k^\top\|_{\mathrm{S}}]$ of the TOPUP estimator for U_k, the initialization of iTOPUP, is no larger than a constant times

$$\begin{aligned} R_k^{(0)} = \lambda_k^{-2}\sigma T^{-1/2}\{&\sqrt{d_kd_{-k}r_{-k}}\big\|\Theta^*_{k,0}\big\|_{\mathrm{S}}^{1/2} + (\sqrt{d_k}+\sqrt{d_{-k}r})\|\Theta_{k,0}\|_{\mathrm{op}}^{1/2} \\ &+ \sigma\sqrt{d_k}d_{-k} + \sigma d_k\sqrt{d_{-k}}T^{-1/2}\}, \end{aligned} \tag{3.7}$$

where $d_{-k} = \prod_{j\neq k} d_j$ and $r_{-k} = \prod_{j\neq k} r_j$. A variation of the Wedin (1972) perturbation theory, stated in Lemma 4.1, provides a sharper bound for the TOPUP estimator as follows.

PROPOSITION 3.1. *Suppose Assumption 1 holds. Let $h_0 \le T/4$. Define*

$$\begin{aligned} \mathscr{R}_{k2} = \lambda_k^{-2}\sigma T^{-1/2}\{&\sqrt{r_kr_{-k}}\big\|\Theta^*_{k,0}\big\|_{\mathrm{S}}^{1/2} + (\sqrt{d_k}+\sqrt{rr_{-k}})\|\Theta_{k,0}\|_{\mathrm{op}}^{1/2} \\ &+ \sigma(\sqrt{d_k}+\sqrt{rr_{-k}}) + \sigma\sqrt{d_kr}T^{-1/2}\}, \end{aligned} \tag{3.8}$$

$$R_k^{(\mathrm{TOPUP})} = \mathscr{R}_{k2} + \big(R_k^{(0)}\big)^2. \tag{3.9}$$

If $\max_{1\le k\le K} R_k^{(0)} = o(1)$, *it holds simultaneously for all* $1\le k\le K$ *that*

$$\overline{\mathbb{E}}\big\|\widehat{P}_k^{(0)} - P_k\big\|_{\mathrm{S}} \lesssim R_k^{(\mathrm{TOPUP})}.$$

REMARK 3.1. In the rank one case ($r_k = 1$, $1 \le k \le K$), Proposition 1 in Ouyang and Yuan (2022) *provides* the sharpness of the above bound. Additionally, for fixed rank, the error bound for TIPUP in Chen, Yang and Zhang (2022b) was confirmed to be sharp by Proposition 1 in Ouyang and Yuan (2022).

The aim of iTOPUP is to achieve dimension reduction by projecting the data in other modes of the tensor time series from $\mathbb{R}^{d_j}$ to $\mathbb{R}^{r_j}$, $j \ne k$. Ideally (e.g., when the true projection matrices U_j are used), this would reduce the rate given in (3.7) and (3.9) to

$$R_k^{(\mathrm{ideal})} = \mathscr{R}_{k2} + \mathscr{R}_{k1}^2, \tag{3.10}$$

by replacing all d_j in $R_k^{(0)}$ with r_j, $j \ne k$, where

$$\begin{aligned}\mathscr{R}_{k1} = \lambda_k^{-2}\sigma T^{-1/2}\{&\sqrt{d_k}r_{-k}\|\Theta_{k,0}^*\|_{\mathrm{S}}^{1/2} + (\sqrt{d_k} + \sqrt{r_{-k}r})\|\Theta_{k,0}\|_{\mathrm{op}}^{1/2} \\ &+ \sigma\sqrt{d_k}r_{-k} + \sigma d_k\sqrt{r_{-k}}T^{-1/2}\}.\end{aligned}$$

However, because the iteration uses the estimated U_j, $j \ne k$, of total dimension $d_{-k}^* = \sum_{j\ne k} d_j r_j$, our analysis also involves the following additional error term:

$$R_k^{(\mathrm{add})} = \lambda_k^{-2}\sigma^2 T^{-1}\Big(d_{-k}^* + \sqrt{d_{-k}^* d_k r_{-k}}\Big). \tag{3.11}$$

The following theorem provides conditions under which the ideal rate is indeed achieved.

THEOREM 3.1. *Suppose Assumption 1 holds. Let* $h_0 \le T/4$ *and* P_k, $\Theta_{k,0}$, $\Theta_{k,0}^*$ *and* λ_k *be as in* (2.2), (3.1), (3.3) *and* (3.5), *respectively. Let* $R^{(0)} = \max_{1\le k\le K} R_k^{(0)}$ *with the* $R_k^{(0)}$ *in* (3.7), $R^{(\mathrm{TOPUP})} = \max_{1\le k\le K} R_k^{(\mathrm{TOPUP})}$ *with the* $R_k^{(\mathrm{TOPUP})}$ *in* (3.9), $R^{(\mathrm{ideal})} = \max_{1\le k\le K} R_k^{(\mathrm{ideal})}$ *with the* $R_k^{(\mathrm{ideal})}$ *in* (3.10) *and* $R^{(\mathrm{add})} = \max_{1\le k\le K} R_k^{(\mathrm{add})}$ *with the* $R_k^{(\mathrm{add})}$ *in* (3.11). *Let* $\widehat{P}_k^{(m)} = \widehat{U}_k^{(m)}\widehat{U}_k^{(m)\top}$ *with the* m*-step estimator* $\widehat{U}_k^{(m)}$ *in the iTOPUP algorithm. Then the following statements hold for a certain numerical constant* $C_1^{(\mathrm{TOPUP})}$ *and a constant* $C_{1,K}^{(\mathrm{iter})}$ *depending on* K *only*: *When*

$$C_1^{(\mathrm{TOPUP})} R^{(0)} \le (1-\rho)/4 \quad \text{and} \quad C_{1,K}^{(\mathrm{iter})}\big(R^{(\mathrm{ideal})} + R^{(\mathrm{add})}\big) \le \rho \tag{3.12}$$

with a constant $0 < \rho < 1$, *it holds simultaneously for all* $1 \le k \le K$ *and* $m \ge 0$ *that*

$$\|\widehat{P}_k^{(m)} - P_k\|_{\mathrm{S}} \le 2C_1^{(\mathrm{TOPUP})}\big((1-\rho^m)(1-\rho)^{-1}R^{(\mathrm{ideal})} + (\rho^m/2)R^{(\mathrm{TOPUP})}\big) \tag{3.13}$$

in an event with probability at least $1 - \sum_{k=1}^K e^{-d_k}$. *In particular, after at most* $J = \lfloor \log(\max_k d_{-k}/r_{-k})/\log(1/\rho)\rfloor$ *iterations*,

$$\overline{\mathbb{E}}\Big[\max_{1\le k\le K}\|\widehat{P}_k^{(J)} - P_k\|_{\mathrm{S}}\Big] \le \frac{3C_1^{(\mathrm{TOPUP})}}{1-\rho}R^{(\mathrm{ideal})} + \sum_{k=1}^K e^{-d_k}. \tag{3.14}$$

REMARK 3.2. The essence of our analysis of iTOPUP is that under (3.12), each iteration is a contraction of the error in the estimation of $\times_{j\ne k}U_j$ in a small neighborhood of it. The upper bound (3.13) for the error of the m-step estimator is comprised of two terms respectively corresponding to the cumulative iteration error and the contracted error of the initial estimator. Of course, after sufficiently large number of iterations, the first term would dominate the second as in (3.14).

REMARK 3.3. The constant $C_1^{(\mathrm{TOPUP})}$ is taken in (3.12) to guarantee sufficient accuracy of the initialization of iTOPUP in the following sense:

$$\max_{k\le K}\overline{\mathbb{E}}\|\widehat{U}_k^{(0)}(\widehat{U}_k^{(0)})^\top - P_k\|_{\mathrm{S}} \le C_1^{(\mathrm{TOPUP})}R^{(0)} \tag{3.15}$$

with at least probability $1-8^{-1}\sum_{k=1}^K e^{-d_k}$. The consistency of the noniterative TOPUP estimator requires $R^{(0)}\to 0$ (Chen, Yang and Zhang (2022a)). However, here we do not require the TOPUP estimator as the initial value to be consistent. For (3.13) to hold, the TOPUP estimator is only required to be sufficiently close to the ground truth as in (3.15).

REMARK 3.4. It is relatively easy to verify that the first part of (3.12) implies the second part under many circumstances, including when d_k are of the same order, r_k are of the same order and $r_k \lesssim d_k^{1-1/K}$ $(K\ge 2)$. In Zhang and Xia (2018), condition $\max_k r_k \lesssim \min_k d_k^{1/2}$ is imposed to control the complexity of the estimated U_j in HOOI although their error bound is sharp and their model is very different. In Corollaries 3.1 and 3.3 below, we prove that the second part of (3.12) follows from the first part respectively in a general fixed rank model and a general diverging rank model. In fact, $R_k^{(\mathrm{ideal})}+R_k^{(\mathrm{add})}\ll R^{(0)}$ typically so that the second part of (3.12) provides a nonasymptotic lower bound for the ρ in (3.13), allowing $\rho=\rho_{T,d_k,d_{-k}^*,r_k,r_{-k},\lambda_k}\to 0$. In Corollary 3.1 below, $\rho = C_{1,K}^{(\mathrm{iter})}(R^{(\mathrm{ideal})}+R^{(\mathrm{add})})$ is taken in (3.12) to give (3.14) in one iteration when $R_k^{(\mathrm{ideal})}$ dominates $R_k^{(\mathrm{add})}$.

REMARK 3.5. When the loading matrices A_k and the TOPUP version of the matrix unfolding of the autocovariance of $\mathcal{F}_t$ all have bounded condition numbers and average squared entries of magnitude 1, λ_k^2, $\|\Theta_{k,0}^*\|_{\mathrm{S}}$ and $\|\Theta_{k,0}\|_{\mathrm{op}}$ are all of the order $d\times \mathrm{poly}(r_1,\dots,r_K)$. In this case, Theorem 3.1 just requires $T\ge \mathrm{poly}(r_1,\dots,r_K)$ for the initialization to achieve through iteration the fast convergence rate $T^{-1/2}d_{-k}^{-1/2}\mathrm{poly}(r_1,\dots,r_K)$; see Corollary 3.3 for details. This is in sharp contrast to the results of traditional factor analysis, which requires $T\to\infty$ to consistently estimate the loading spaces. The main reason is that the other tensor modes provide additional information and in certain sense serve as additional samples. Roughly speaking, we have totally $dT=d_kd_{-k}T$ observations in the tensor time series to estimate the d_kr_k parameters in the projection to the column space of the loading matrix A_k, where $r_k\ll d_{-k}T$ in the above "regular" case.

Now, let us consider the statistical performance of iTIPUP procedure. Again, by Chen, Yang and Zhang (2022a) the TIPUP risk in the estimation of P_k is bounded by

$$\overline{\mathbb{E}}[\|\widehat{P}_k^{(\mathrm{TIPUP})}-P_k\|_{\mathrm{S}}] \lesssim R_k^{*(0)} = (\lambda_k^*)^{-2}\sigma T^{-1/2}\sqrt{d_k}\big(\|\Theta_{k,0}^*\|_{\mathrm{S}}^{1/2}+\sigma\sqrt{d_{-k}}\big) \tag{3.16}$$

with $d_{-k}=\prod_{j\ne k}d_j$, and the aim of iTIPUP is to achieve the ideal rate

$$R_k^{*(\mathrm{ideal})} = (\lambda_k^*)^{-2}\sigma T^{-1/2}\sqrt{d_k}\big(\|\Theta_{k,0}^*\|_{\mathrm{S}}^{1/2}+\sigma\sqrt{r_{-k}}\big) \tag{3.17}$$

through dimension reduction, where $r_{-k}=\prod_{j\ne k}r_j$. As in the case of iTOPUP, our error bound for iTIPUP involves the additional error term

$$R_k^{*(\mathrm{add})} = \sqrt{d_{-k}^*/d_k}\,R_k^{*(\mathrm{ideal})}. \tag{3.18}$$

The following theorem, which allows the ranks r_k to grow to infinity as well as d_k when $T\to\infty$, provides sufficient conditions to guarantee the ideal convergence rate for iTIPUP.

THEOREM 3.2. *Suppose Assumption 1 holds. Let* P_k, $\Theta^*_{k,0}$ *and* λ^*_k *be as in* (2.2), (3.3) *and* (3.6), *respectively. Let* $h_0 \le T/4$, *and*

$$R^{*(0)} = \max_{1\le k\le K} R_k^{*(0)}; \qquad R^{*(\mathrm{ideal})} = \max_{1\le k\le K} R_k^{*(\mathrm{ideal})}, \quad R^{*(\mathrm{add})} = \max_{1\le k\le K} R_k^{*(\mathrm{add})}$$

with $R_k^{*(0)}$ *in* (3.16), $R_k^{*(\mathrm{ideal})}$ *in* (3.17) *and* $R_k^{*(\mathrm{add})}$ *in* (3.18). *Let* $\widehat{P}_k^{(m)} = \widehat{U}_k^{(m)}\widehat{U}_k^{(m)\top}$ *with the* m*-step estimator* $\widehat{U}_k^{(m)}$ *in the iTIPUP algorithm. Then the following statements hold for a certain numerical constant* $C_1^{(\mathrm{TIPUP})}$ *and a constant* $C_{1,K}^{(\mathrm{iter})}$ *depending on* K *only*: *When*

$$C_1^{(\mathrm{TIPUP})} R^{*(0)} \le \min_{1\le k\le K} \frac{(1-\rho)\lambda_k^{*2}}{8\|\Theta^*_{k,0}\|_{\mathrm{S}}} \quad \text{and} \quad C_{1,K}^{(\mathrm{iter})}\big(R^{*(\mathrm{ideal})} + R^{*(\mathrm{add})}\big) \le \rho \tag{3.19}$$

with a constant $0 < \rho < 1$, *it holds simultaneously for all* $1 \le k \le K$ *and* $m \ge 0$ *that*

$$\big\|\widehat{P}_k^{(m)} - P_k\big\|_{\mathrm{S}} \le 2C_1^{(\mathrm{TIPUP})}\big((1-\rho^m)(1-\rho)^{-1}R^{*(\mathrm{ideal})} + (\rho^m/2)R^{*(0)}\big) \tag{3.20}$$

in an event with probability at least $1 - \sum_{k=1}^K e^{-d_k}$. *In particular, after at most* $J = \lfloor \log(\max_k d_{-k}/r_{-k})/\log(1/\rho)\rfloor$ *iterations*,

$$\overline{\mathbb{E}}\Big[\max_{1\le k\le K}\big\|\widehat{P}_k^{(J)} - P_k\big\|_{\mathrm{S}}\Big] \le \frac{3C_1^{(\mathrm{TIPUP})}}{1-\rho}R^{*(\mathrm{ideal})} + \sum_{k=1}^K e^{-d_k}. \tag{3.21}$$

We briefly discuss the conditions and conclusions of Theorem 3.2 as the details are parallel to the remarks below Theorem 3.1. By (3.3), (3.6) and the Cauchy–Schwarz inequality, $(1-h_0/T)\lambda_k^{*2} \le \|\Theta^*_{k,0}\|_{\mathrm{S}}$, so that the first condition in (3.19) guarantees a sufficiently small $R^{*(0)}$, which implies a sufficiently small error in the initialization of iTIPUP by (3.16). The second condition in (3.19) again has two terms respectively reflecting the ideal rate after dimension reduction by the true $U_{-k} = \odot_{j\ne k}U_j$ in the estimation of U_k and the extra cost of estimating U_{-k}. The upper bound (3.20) for the error of the m-step estimator is also comprised of two terms representing the cumulative iteration error and contracted initialization error. In Corollary 3.2 below with fixed r_k, the smallest $\rho = C_{1,K}^{(\mathrm{iter})}(R^{*(\mathrm{ideal})} + R^{*(\mathrm{add})})$ is taken in (3.19) to achieve (3.21) in one iteration when $R_k^{*(\mathrm{ideal})}$ dominates $R_k^{*(\mathrm{add})}$. Moreover, Theorem 3.2 allows diverging ranks r_k and convergence rate $T^{-1/2}d_{-k}^{-1/2}\mathrm{poly}(r_1,\dots,r_K)$ under proper conditions as discussed in Remark 3.5.

As discussed in Section 2.3, we can mix the TOPUP and TIPUP operations for the initiation and iterative operations in Algorithm 1. For example, the proof of Theorems 3.1 yields the following error bound for the mixed TIPUP-iTOPUP algorithm.

THEOREM 3.3. *Assumption 1 holds. Let* $R^{(0)}$, $R^{(\mathrm{ideal})}$ *and* $R^{(\mathrm{add})}$ *be as in Theorem 3.1 and* $R^{*(0)}$ *be as in Theorem 3.2. Let* $\widehat{P}_k^{(m)} = \widehat{U}_k^{(m)}\widehat{U}_k^{(m)\top}$ *with* $\widehat{U}_k^{(m)}$ *being the* m*-step estimator in the* TIPUP-iTOPUP *algorithm. Then the following statement holds for a certain numerical constant* $C_1^{(\mathrm{TOPUP})}$ *and a constant* $C_{1,K}^{(\mathrm{iter})}$ *depending on* K *only*: *When*

$$C_1^{(\mathrm{TOPUP})} R^{*(0)} \le (1-\rho)/4 \quad \text{and} \quad C_{1,K}^{(\mathrm{iter})}\big(R^{(\mathrm{ideal})} + R^{(\mathrm{add})}\big) \le \rho \tag{3.22}$$

with a constant $0 < \rho < 1$, *it holds in an event with probability at least* $1 - \sum_{k=1}^K e^{-d_k}$ *that simultaneously for all* $1 \le k \le K$ *and* $m \ge 0$

$$\big\|\widehat{P}_k^{(m)} - P_k\big\|_{\mathrm{S}} \le 2C_1^{(\mathrm{TOPUP})}\big((1-\rho^m)(1-\rho)^{-1}R^{(\mathrm{ideal})} + (\rho^m/2)R^{*(0)}\big).$$

We omit the statement of an analogous error bound for the TOPUP-iTIPUP algorithm.

3.3. *Fixed rank factor process.* In this section, we provide the convergence rate when the dimensions of the factors $\mathcal{F}_t$, or equivalently the ranks of the signal process $\mathcal{M}_t$, $r_1, \ldots, r_K$, are fixed, and the autocross-outer product of the factor process is ergodic. Formally, we impose the following additional assumption.

ASSUMPTION 2. The ranks $r_1, \ldots, r_K$ are fixed. The factor process $\mathcal{F}_t$ is weakly stationary and its autocross-outer-product process is ergodic in the sense of

$$\frac{1}{T-h}\sum_{t=h+1}^{T} \mathcal{F}_{t-h} \otimes \mathcal{F}_t \longrightarrow \mathbb{E}(\mathcal{F}_{t-h} \otimes \mathcal{F}_t) \quad \text{in probability,}$$

where the elements of $\mathbb{E}(\mathcal{F}_{t-h} \otimes \mathcal{F}_t)$ are all finite. In addition, the condition numbers of $A_k^\top A_k$ $(k = 1, \ldots, K)$ are bounded. Furthermore, assume that h_0 is fixed, and:

(i) (TOPUP related): $\mathbb{E}[\mathrm{mat}_1(\Phi_{k,1:h_0})]$ is of rank r_k for $1 \le k \le K$.

(ii) (TIPUP related): $\mathbb{E}[\Phi^{*(\mathrm{cano})}_{k,1:h_0}]$ is of rank r_k for $1 \le k \le K$.

Under Assumption 2, the factor process has a fixed expected autocross-moment tensor with fixed dimensions. The assumption that the condition numbers of $A_k^\top A_k$ $(k = 1, \ldots, K)$ are bounded corresponds to the pervasive condition (e.g., Stock and Watson (2002), Bai (2003)). It ensures that all the singular values of A_k are of the same order. Such conditions are commonly imposed in factor analysis.

As our methods are based on autocross-moment at nonzero lags, we do not need to assume any specific model for the latent process $\mathcal{F}_t$, except some rank conditions in Assumption 2(i) and (ii). Since the columns of $\Phi^{*(\mathrm{cano})}_{k,1:h_0}$ are linear combinations of those of $\mathrm{mat}_1(\Phi^{(\mathrm{cano})}_{k,1:h_0})$ and $\mathbb{E}[\mathrm{mat}_1(\Phi_{k,1:h_0})]$ and $\mathbb{E}[\mathrm{mat}_1(\Phi^{(\mathrm{cano})}_{k,1:h_0})]$ have the same rank, Assumption 2(ii) implies Assumption 2(i).

In order to provide a more concrete understanding of Assumption 2(i) and (ii), consider the case of $k = 1$ and $K = 2$. We write the factor process $\mathcal{F}_t = (f_{i,j,t})_{r_1 \times r_2}$, and the stationary autocross-moments $\phi_{i_1,j_1,i_2,j_2,h} = \mathbb{E}(f_{i_1,j_1,t-h} f_{i_2,j_2,t})$. Hence, $\mathbb{E}[\mathrm{mat}_1(\Phi_{k,1:h_0})]$ is a $r_k \times (r_{-k} r_k r_{-k} h_0)$ matrix, with columns being $\phi_{\cdot,j_1,i_2,j_2,h}$. Since $\mathbb{E}[\mathrm{mat}_1(\Phi_{k,1:h_0})] \times \mathbb{E}[\mathrm{mat}_1(\Phi_{k,1:h_0})]^\top$ is a sum of many semipositive definite $r_k \times r_k$ matrices, if any one of these matrices is full rank, then $\mathbb{E}[\mathrm{mat}_1(\Phi_{k,1:h_0})]$ is of rank r_k. Hence, Assumption 2(i) is relatively easy to fulfill. On the other hand, Assumption 2(ii) is quite different. First, the condition is imposed on the canonical form of the model as the inner product in TIPUP related procedures behaves differently. Let $\mathcal{F}_t^{(\mathrm{cano})} = U_1^\top \mathcal{M}_t U_2 = (f^{(\mathrm{cano})}_{i,j,t})_{r_1 \times r_2}$, and $\phi^{(\mathrm{cano})}_{i_1,j_1,i_2,j_2,h} = \mathbb{E}(f^{(\mathrm{cano})}_{i_1,j_1,t-h} f^{(\mathrm{cano})}_{i_2,j_2,t})$. Then $\|\Phi^{*(\mathrm{cano})}_{1,1:h_0}\|_{\mathrm{HS}}^2 = \sum_{h=1}^{h_0} \sum_{i_1,i_2} (\sum_{j=1}^{r_2} \phi^{(\mathrm{cano})}_{i_1,j,i_2,j,h})^2$. As $\phi^{(\mathrm{cano})}_{i_1,j,i_2,j,h}$ may be positive or negative for different i_1, i_2, j, h, the summation $\sum_{j=1}^{r_2} \phi^{(\mathrm{cano})}_{i_1,j,i_2,j,h}$ is subject to potential signal cancellation for $h > 0$. Assumption 2(ii) ensures that there is no complete signal cancellation that makes the rank of $\mathbb{E}[\Phi^{*(\mathrm{cano})}_{k,1:h_0}]$ less than r_k. While the signal cancellation rarely causes the rank deficiency, the resulting loss of efficiency may still have an impact on the finite sample performance as our simulation results demonstrate. Of course, complete signal cancellation is less likely with larger h_0.

The following corollary is a simplified version of Theorem 3.1 under Assumption 2(i).

COROLLARY 3.1. *Suppose Assumptions 1 and 2(i) hold. Let* $\lambda = \prod_{k=1}^K \|A_k\|_{\mathrm{S}}$ *and* $r_{-k} = r/r_k$. *Let* $h_0 \le T/4$ *and* σ *fixed. Then there exist numerical constants* $C_{0,K}$ *and* $C_{1,K}$ *depending on* K *only such that when*

$$\lambda^2 \ge C_{0,K}\sigma^2 \max_{1\le k\le K}\left(\frac{d r_{-k}}{T} + \frac{d}{\sqrt{T d_k r_{-k}}}\right), \tag{3.23}$$

the 1*-step iTOPUP estimator satisfies*

$$\mathbb{E}\|\widehat{P}_k^{(1)} - P_k\|_{\mathrm{S}} \le C_{1,K}\left(\frac{\sigma}{\lambda\sqrt{T}}\left(\frac{\sqrt{d_k}}{\sqrt{r_{-k}}} + \sqrt{r r_{-k}}\right) + \frac{\sigma^2}{\lambda^2\sqrt{T}}\left(\frac{\sqrt{d_k}}{\sqrt{r_{-k}}} + \sqrt{r}\right)\right) + C_{1,K}\left(\frac{\sigma\sqrt{d_k r_{-k}}}{\lambda\sqrt{T}} + \frac{\sigma^2\sqrt{d_k r_{-k}}}{\lambda^2\sqrt{T}}\right)^2 + \sum_{k=1}^{K} e^{-d_k}. \tag{3.24}$$

REMARK 3.6. Under Assumption 2 that r_k is fixed, (3.23), (3.24) and (3.25), (3.26) in Corollary 3.2 can absorb r_k's into the numerical constants. The corollaries are expressed in this form to allow divergent r_k for the purpose of facilitating comparison with the minimax lower bound in Theorem 3.5. They also represent a specific case of Corollaries 3.3–3.5.

Corollary 3.1 asserts that, in order to recover the factor loading space for A_k, the signal-to-noise ratio needs to satisfy $\lambda/\sigma \ge C_{0,K,r}\max_{k\le K}(d^{1/2}T^{-1/2} + d^{1/2}d_k^{-1/4}T^{-1/4})$ as in (3.23), and the ideal rate (3.24) can be achieved in one iteration. Under Assumptions 1, 2(i), the error bound in Proposition 3.1 yields the convergence rate

$$\mathbb{E}\|\widehat{P}_k^{(0)} - P_k\|_{\mathrm{S}} \lesssim \left(\frac{\sigma\sqrt{d_k}}{\lambda\sqrt{T}} + \frac{\sigma^2\sqrt{d_k}}{\lambda^2\sqrt{T}}\right) + \left(\frac{\sigma\sqrt{d}}{\lambda\sqrt{T}} + \frac{\sigma^2\sqrt{d_k}d_{-k}}{\lambda^2\sqrt{T}}\right)^2.$$

In comparison, the ideal rate is much sharper than the convergence rate of the noniterative TOPUP in Proposition 3.1 when $\lambda^2/\sigma^2 \ll \min_{k\le K}\{d^{4/3}/(T^{1/3}d_k), d^2/(T^{1/2}d_k^{3/2})\}$.

The following corollary is a simplified version of Theorem 3.2 under Assumption 2(ii), which excludes severe signal cancellation in iTIPUP.

COROLLARY 3.2. *Suppose Assumptions* 1 *and* 2(ii) *hold. Let* $\lambda = \prod_{k=1}^{K}\|A_k\|_{\mathrm{S}}$ *and* $r_{-k} = r/r_k$. *Let* $h_0 \le T/4$ *and* σ *fixed. Then there exist constants* $C_{0,K}$ *and* $C_{1,K}$ *depending on* K *only such that when*

$$\lambda^2 \ge C_{0,K}\sigma^2 \max_{1\le k\le K}\left(\frac{d_k}{T r_{-k}} + \frac{\sqrt{d}}{\sqrt{T}r_{-k}}\right), \tag{3.25}$$

the 1*-step iTIPUP estimator satisfies*

$$\mathbb{E}\|\widehat{P}_k^{(1)} - P_k\|_{\mathrm{S}} \le C_{1,K}\left(\frac{\sigma\sqrt{d_k}}{\lambda\sqrt{T r_{-k}}} + \frac{\sigma^2\sqrt{d_k}}{\lambda^2\sqrt{T r_{-k}}}\right) + \sum_{k=1}^{K} e^{-d_k}, \tag{3.26}$$

and the 1*-step TIPUP-iTOPUP estimator satisfies* (3.24).

Compared with the results in Corollary 3.1 for iTOPUP, the achieved ideal rate (3.26) is the same. However, the signal-to-noise ratio requirement (3.25) is weaker but Assumption 2(ii) is stronger in Corollary 3.2 for iTIPUP. Again, the ideal rate is much sharper than the convergence rate of the noniterative TIPUP in Chen, Yang and Zhang (2022a).

3.4. *Diverging ranks.* The main theorems in Section 3.2 allow for the case where the dimensions of the core factor, $r_1, \ldots, r_K$, diverge as the dimensions of the observed tensor $d_1, \ldots, d_K$ grow to infinity. The following assumption provides a concrete set of conditions that can be used to provide some insights of the properties of iTOPUP and iTIPUP in such scenarios.

ASSUMPTION 3. For a certain $\delta_0 \in [0, 1]$, $\|\Theta_{k,0}\|_{\mathrm{op}} \asymp \sigma^2 d^{1-\delta_0}/r$ and $\|\Theta_{k,0}^*\|_{\mathrm{S}} \asymp \sigma^2 d^{1-\delta_0}/r_k$ with probability approaching one. For the singular values, two scenarios are considered.

(i) (TOPUP related): There exist some constants $\delta_1 \in [\delta_0, 1]$ and $c_1 > 0$ such that with probability approaching one (as $T \to \infty$) $\lambda_k^2 \geq c_1 \sigma^2 d^{1-\delta_1}/\sqrt{r r_k}$, for all $k = 1, \ldots, K$.

(ii) (TIPUP related): There exist some constants $\delta_1 \in [\delta_0, 1]$, $c_2 > 0$ and $\delta_2 \geq 0$ such that with probability approaching one (as $T \to \infty$), $\lambda_k^{*2} \geq c_2 \sigma^2 d^{1-\delta_1} r_k^{-1} r_{-k}^{-\delta_2}$ for all $k = 1, \ldots, K$.

Assumption 3 is similar to the signal strength condition of Lam and Yao (2012), and the pervasive condition on the factor loadings (e.g., Stock and Watson (2002) and Bai (2003)). It is more general than Assumption 2 in the sense that it allows $r_1, \ldots, r_K$ to diverge and the latent process $\mathcal{F}_t$ does not have to be weakly stationary.

We take δ_0, δ_1 as measures of the strength of the signal process $\mathcal{M}_t$. They roughly indicate how much information is contained in the signals compared with the amount of noise, with respect to the dimensions and ranks, d, r and r_k. In this sense, they reflect the signal-to-noise ratio. When $\delta_0 = \delta_1 = 0$, the factors are called strong factors; otherwise, the factors are called weak factors.

REMARK 3.7 (Signal strength and the index δ_0). We note that $\operatorname{trace}(\Theta_{k,0}) = \operatorname{trace}(\Theta_{k,0}^*) = \sum_{t=1}^T \|\operatorname{vec}(\mathcal{M}_t)\|_2^2/T$, and that $\operatorname{rank}(\Theta_{k,0}) = r$ and $\operatorname{rank}(\Theta_{k,0}^*) = r_k$ when the data is in general position, where $\Theta_{k,0}$ is treated as a $d \times d$ matrix. Thus, if $\sum_{t=1}^T \|\operatorname{vec}(\mathcal{M}_t)\|_2^2/(\sigma^2 dT) \asymp d^{-\delta_0}$ is the signal-to-noise ratio, then the condition $\|\Theta_{k,0}\|_{\mathrm{op}} \asymp \sigma^2 d^{1-\delta_0}/r$ holds when r is the order of the effective rank of $\Theta_{k,0}$ and the condition $\|\Theta_{k,0}^*\|_{\mathrm{S}} \asymp \sigma^2 d^{1-\delta_0}/r_k$ holds when r_k is the order of the effective rank of $\Theta_{k,0}^*$. Because the signal $\mathcal{M}_t$ has d elements at each t, the assumption $\sum_{t=1}^T \|\operatorname{vec}(\mathcal{M}_t)\|_2^2/(\sigma^2 dT) \asymp d^{-\delta_0}$ says that the squared ratio of the elements and the noise level is $d^{-\delta_0}$ averaged over time and space. Thus, the factor is called strong when $\delta_0 = 0$. In view of (1.1) and (1.2), $\mathcal{M}_t = \mathcal{F}_t \times_{k=1}^K A_k$, so that we may have weaker factor with $\delta_0 > 0$ when the loading matrices A_k are sparse or have some relatively small singular components. We note that by Cauchy–Schwarz, the signal-to-noise ratio conditions also imply $(1 - h/T)^2 \|\Theta_{k,h}\|_{\mathrm{HS}}^2 \leq \|\Theta_{k,0}\|_{\mathrm{HS}}^2 \lesssim r(\sigma^2 d^{1-\delta_0}/r)^2$ and $(1 - h/T)^2 \|\Theta_{k,h}^*\|_{\mathrm{HS}}^2 \leq \|\Theta_{k,0}^*\|_{\mathrm{HS}}^2 \lesssim r_k(\sigma^2 d^{1-\delta_0}/r_k)^2$, respectively.

REMARK 3.8 (Assumption 3(i) and the role of δ_1). In fact, for TOPUP, Assumption 3(i) holds when (a) $\|\overline{\mathbb{E}}[\mathrm{TOPUP}_k]\|_{\mathrm{HS}}^2 = \sum_{h=1}^{h_0} \|\Theta_{k,h}\|_{\mathrm{HS}}^2 \asymp h_0 \sigma^4 d^{2(1-\delta_1)}/r$ and (b) all the nonzero singular values of $\overline{\mathbb{E}}[\mathrm{TOPUP}_k]$ are of the same order. Because $\|\Theta_{k,h}\|_{\mathrm{HS}}^2 \lesssim \sigma^4 d^{2(1-\delta_0)}/r$ by the condition on the signal-to-noise ratio, we must have $\delta_1 \geq \delta_0$, and $d^{\delta_0 - \delta_1}$ can be viewed as the order of average autocorrelation over lags $h = 1, \ldots, h_0$. For $k = 1$ and $K = 2$, the factor process in the canonical form is $\mathcal{F}_t^{(\mathrm{cano})} = U_1^\top \mathcal{M}_t U_2 = (f_{i,j,t}^{(\mathrm{cano})})_{r_1 \times r_2}$, and $\phi_{i_1,j_1,i_2,j_2,h}^{(\mathrm{cano})} = \sum_{t=h+1}^T f_{i_1,j_1,t-h}^{(\mathrm{cano})} f_{i_2,j_2,t}^{(\mathrm{cano})}/(T-h)$ is the time average cross-product between the factor fibers $f_{i_1,j_1,1:T}^{(\mathrm{cano})}$ and $f_{i_2,j_2,1:T}^{(\mathrm{cano})}$. Thus, the first condition (a) means $\sum_{h=1}^{h_0} \|\Theta_{1,h}\|_{\mathrm{HS}}^2 = \sum_{h=1}^{h_0} \|\Phi_{1,h}^{(\mathrm{cano})}\|_{\mathrm{HS}}^2 = \sum_{i_1,j_1,i_2,j_2,h} (\phi_{i_1,j_1,i_2,j_2,h}^{(\mathrm{cano})})^2 \asymp h_0 \sigma^4 d^{2(1-\delta_1)}/r$.

REMARK 3.9 (Assumption 3(ii), the role of δ_2 and signal cancellation). The points parallel to those in Remark 3.8 are applicable to TIPUP, but with one caveat: Beyond the average autocorrelation, an additional discount $r_{-k}^{-\delta_2} \leq 1$ is needed to take into account the impact of possible signal cancellation with TIPUP and its iteration. For $k = 1$ and $K = 2$, $\|\Theta_{1,h}^*\|_{\mathrm{HS}}^2 = \|\Phi_{1,h}^{*(\mathrm{cano})}\|_{\mathrm{HS}}^2 = \sum_{i_1,i_2} (\sum_{j=1}^{r_2} \phi_{i_1,j,i_2,j,h}^{(\mathrm{cano})})^2$ and the summation inside the square is subject to signal cancellation for $h > 0$ since the autocross-moment $\phi_{i_1,j,i_2,j,h}^{(\mathrm{cano})}$ can have different signs. The additional parameter δ_2 measures the severity of signal cancellation in the TIPUP related procedures. For example, when the majority of $\phi_{i_1,j,i_2,j,h}^{(\mathrm{cano})}$ are of the same

sign for most of (i_1, i_2, h), it would be reasonable to assume $\delta_2 = 0$. When $\phi^{(\mathrm{cano})}_{i_1,j,i_2,j,h}$ behave like independent mean zero variables, δ_2 would be close to 0.5. And $\delta_2 = \infty$ when all the signals cancel out by the summation $\phi^{(\mathrm{cano})}_{i_1,j,i_2,j,h}$ over j. In the case of fixed r_k, the convergence rate depends on whether $\delta_2 = \infty$ (severe signal cancellation) or not.

REMARK 3.10 (The role of h_0). The selection of h_0 is a relative minor problem in practice though very complex to analyze. Theoretically, it suffices to use an h_0 with λ_k of the right order, so that choosing a somewhat large h_0 would not harm the convergence rate for the proposed methods. In practice, a small h_0 (less than 3) is often sufficient. The impact of the choice of h_0 on the signal and noise depends on the autocorrelation of the factor process, as well as the loading matrices. For example, if the factor process is of very short memory (e.g., an MA(1) process), including any lag $h > 1$ only introduces noise to TOPUP_k in (2.4) and TIPUP_k in (2.7) without enhancing the signal. On the other hand, including an extra lag is the most simple and effective way to prevent signal cancellation with iTIPUP, as discussed in the previous remark. Increasing h_0 includes more nonnegative terms in the signal strength $\sum_{i_1,i_2,h}(\sum_{j=1}^{r_2}\phi^{(\mathrm{cano})}_{i_1,j,i_2,j,h})^2$, hence potentially reducing the chance of severe signal cancellation. The simulation results presented in the Supplementary Material provide some empirical behavior of choosing different h_0. While the choice of h_0 will affect the assumptions, in practice we may compare the patterns of estimated singular values under different lag values h_0 in iTOPUP and iTIPUP to evaluate the benefit of taking a larger h_0; see also the simulation study.

We describe below the convergence rate of iTOPUP in terms of d_k, r_k and T under Assumption 3(i) when the dimensions of the core factor $r_1, \ldots, r_K$ are allowed to diverge.

COROLLARY 3.3. *Suppose Assumptions 1 and 3(i) hold. Let* $h_0 \le T/4$, $d^*_{-k} = \sum_{j\neq k} d_j r_j$ *and* $r = \Pi_{k=1}^K r_k$. *Suppose that for a sufficiently large* C_0 *not depending on* $\{\sigma, d_k, r_k, k \le K\}$,

$$T \ge C_0 \max_{1\le k\le K}\big(d^{2\delta_1-\delta_0} r_k r^2_{-k} + d^{2\delta_1} r_k^2 r_{-k}/d_k\big). \tag{3.27}$$

Then, after $J = O(\log d)$ *iterations, we have the following upper bounds for iTOPUP*:

$$\begin{aligned}&\max_{1\le k\le K}\big\|\widehat{P}_k^{(J)} - P_k\big\|_{\mathrm{S}}\\ &= O_{\mathbb{P}}(1)\max_{1\le k\le K}\Bigg(\frac{d_k^{1/2} r_k^{1/2}(1 + r^{1/2}/d^{(1-\delta_0)/2}) + r^{3/2} r_k^{-1/2}(1 + r_k^{1/2}/d^{(1-\delta_0)/2})}{T^{1/2} d^{1/2+\delta_0/2-\delta_1}}\\ &\quad + \bigg(\frac{d_k^{1/2} r^{3/2}(1 + r_k^{1/2}/d^{(1-\delta_0)/2})}{T^{1/2} d^{1/2+\delta_0/2-\delta_1} r_k}\bigg)^2\Bigg).\end{aligned} \tag{3.28}$$

Moreover, (3.28) *holds after at most* $J = O(\log r)$ *iterations, if any one of the following three conditions holds in addition to* (3.27): (i) d_k $(k = 1, \ldots, K)$ *are of the same order,* (ii) λ_k $(k = 1, \ldots, K)$ *are of the same order,* (iii) $(\lambda_k)^{-2}\sqrt{d_k}$ $(k = 1, \ldots, K)$ *are of the same order.*

Note that the second part of Corollary 3.3 says that when the condition is right, iTOPUP algorithm only needs a small number of iterations to converge, as $O(\log r)$ is typically very small. The noise level σ does not appear directly in the rate since it is incorporated in the signal-to-noise ratio in the tensor form in Assumption 3. In Corollary 3.3, we show that as long as the sample size T satisfies (3.27), the iTOPUP achieves consistent estimation under proper regularity conditions. To digest the condition, we notice that (3.27) becomes

$T \geq C_0 \max_k(r_k r_{-k}^2)$ when the growth rate of r_k is much slower than d_k and the factors are strong with $\delta_0 = \delta_1 = 0$.

The advantage of using index δ_0, δ_1 is to link the convergence rates of the estimated factor loading space explicitly to the strength of factors. It is clear that the stronger the factors are, the faster the convergence rate is. Equivalently, the stronger the factors are, the smaller the sample size is required.

When the ranks r_k $(k = 1, \ldots, K)$ also diverge and there is no severe signal cancellation in iTIPUP, we have the following convergence rate for iTIPUP under Assumption 3(ii).

COROLLARY 3.4. *Suppose Assumptions 1 and 3(ii) hold. Let* $h_0 \leq T/4$ *and* $d_{-k}^* = \sum_{j \neq k} d_j r_j$. *Suppose that for a sufficiently large* C_0 *not depending on* $\{\sigma, d_k, r_k, k \leq K\}$,

$$T \geq C_0 \max_{1 \leq k \leq K} \left(\frac{(d_k r_k + d^{\delta_0} r_k^2) r_{-k}^{2\delta_2} r^{2\delta_2}}{d^{1+3\delta_0-4\delta_1} \min_{1 \leq k \leq K} r_k^{2\delta_2}} + \frac{d_{-k}^* r_k r_{-k}^{2\delta_2}}{d^{1+\delta_0-2\delta_1}} \left(1 + \frac{r}{d^{1-\delta_0}}\right)\right). \tag{3.29}$$

Then, after at most $J = O(\log d)$ *iterations, the iTIPUP estimator satisfies*

$$\max_{1 \leq k \leq K} \|\widehat{P}_k^{(J)} - P_k\|_{\mathrm{S}} = O_{\mathbb{P}}(1) \max_{1 \leq k \leq K} \left(\frac{d_k^{1/2} r_k^{1/2} r_{-k}^{\delta_2} (1 + r^{1/2}/d^{(1-\delta_0)/2})}{T^{1/2} d^{1/2+\delta_0/2-\delta_1}} \right). \tag{3.30}$$

Moreover, (3.30) *holds after at most* $J = O(\log r)$ *iterations, if any one of the following three conditions holds in addition to condition* (3.29): (i) d_k $(k = 1, \ldots, K)$ *are of the same order,* (ii) λ_k^* $(k = 1, \ldots, K)$ *are of the same order and* (iii) $(\lambda_k^*)^{-2}\sqrt{d_k}$ $(k = 1, \ldots, K)$ *are of the same order.*

When the average autocorrelation is of unit order and the signal cancellation for TIPUP has no impact on the order of the signal ($\delta_0 = \delta_1$ and $\delta_2 = 0$, respectively), Corollary 3.4 requires the sampling rate $T \gtrsim h_0 + (d_k r_k + d^{\delta_0} r_k^2 + d_{-k}^* r_k (1 + r/d^{1-\delta_0}))/d^{1-\delta_0}$ and provides the convergence rate $(r_k d_k)^{1/2}(1 + r/d^{1-\delta_0})^{1/2}/(T d^{1-\delta_0})^{1/2}$. For examples, $T \geq 4h_0 + C_1$ gives the rate $(r_k d_k)^{1/2}/(T d^{1-\delta_0})^{1/2}$ when $\delta_0 \leq (K-2)/(2K)$ and $r_k^2 \lesssim d_k \asymp d^{1/K}$ $\forall k$, and the sample size requirement can be written as $T \gtrsim h_0 + d^{\delta_0} r_k^2/d^{1-\delta_0}$ when $r_k^2 \asymp r^{2/K} \lesssim d_k \asymp d^{1/K}$ $\forall k$ regardless of $\delta_0 \in [1/K, 1]$. Thus, the side condition involving $R^{*(\mathrm{add})}$ in the second part of (3.19) is absorbed into the other components of (3.19).

Corollary 3.3 and Corollary 3.4 offer comparison of the iTOPUP and iTIPUP when the ranks diverge from two perspectives: sample size requirements and convergence rates. The lower bounds on T in (3.27) in Corollary 3.3 and (3.29) in Corollary 3.4 provide the sample complexity of the iTOPUP and iTIPUP, respectively. In the case that the growth rate of r_k is much slower than d_k and the factors are strong with $\delta_0 = \delta_1 = 0$, the required sample size of the iTIPUP reduces to $T \geq 4h_0 + C_0 \max_{j,k}(r_k r_{-k}^{2\delta_2} r_{-j}^{2\delta_2}/d_{-k} + r_k r_{-k}^{2\delta_2} r_j/d_{-j})$, where $r_{-k} = r/r_k$ and $d_{-k} = d/d_k$. By comparing with the comment after Corollary 3.3, where the sample size requirement for the iTOPUP is $T \geq C_0 \max_k(r_k r_{-k}^2)$ when $\delta_0 = \delta_1 = 0$, it can be seen that the sample complexity for the iTIPUP is smaller, if δ_2 is a small constant. From the perspective of convergence rate, let us compare (3.28) in Corollary 3.3 and (3.30) in Corollary 3.4. When ranks diverge, iTIPUP is slower than iTOPUP if $\delta_2 > 3/2$, or $\{0 \leq \delta \leq 3/2, d_k \gtrsim r r_{-k}^{2-2\delta_2}, d_{-k} \gtrsim r r_{-k}^{3-2\delta_2}\}$, and faster if $d_k \lesssim r_k r_{-k}^{2-2\delta_2}$, no matter how strong the factor is or what values δ_0, δ_1 take. As expected, the convergence rate is slower in the presence of weak factors; see the simulation for more empirical evidence.

Similar to Corollaries 3.3 and 3.4, we have the following rate for TIPUP-iTOPUP.

COROLLARY 3.5. *Suppose Assumptions 1 and 3 hold. Let* $h_0 \le T/4$ *and* $d_{-k}^* = \sum_{j\neq k} d_j r_j$. *Suppose that for a sufficiently large* C_0 *not depending on* $\{\sigma, d_k, r_k, k \le K\}$,

(3.31)
$$T \ge C_0 \max_{1\le k\le K}\left(d^{2\delta_1-\delta_0} r_k\left(\frac{r_{-k}^{2\delta_2}}{d_{-k}} + \frac{r_{-k}^3}{d_{-k}}\right) + \frac{d^{2\delta_1} r_k^2}{d_k}\left(\frac{r_{-k}^{2\delta_2}}{d_{-k}} + \frac{r_{-k}^3}{d_{-k}^2}\right) + \frac{d_{-k}^*\sqrt{r r_k}}{d^{1-\delta_1}}\right).$$

Then, after at most $J = O(\log d)$ *iterations, the TIPUP-iTOPUP estimator satisfies* (3.28). *Moreover, the above error bound holds after at most* $J = O(\log r)$ *iterations, if any one of the following three conditions holds in addition to condition* (3.31): (i) d_k $(k = 1, \ldots, K)$ *are of the same order,* (ii) λ_k $(k = 1, \ldots, K)$ *are of the same order and* (iii) $(\lambda_k)^{-2}\sqrt{d_k}$ $(k = 1, \ldots, K)$ *are of the same order.*

Compared with Corollary 3.3, Corollary 3.5 provides the same error bound for smaller T (possibly with bounded $T \gtrsim h_0$) when $r_{-k}^{2\delta_2} \lesssim r_{-k} d_{-k}$. The side condition involving $R^{(\mathrm{add})}$ in the second part of (3.22), corresponding to the last component of (3.31) involving d_{-k}^*, is absorbed into the other components of (3.22) when $r_k^{1/2} \le d^{\delta_1-\delta_0}(r_{-k}^{2\delta_2-1} + r_{-k}^2)$ $\forall k \le K$.

3.5. *Comparisons.*

3.5.1. *Comparison between the noniterative procedures and iterative procedures.* Theorems 3.1 and 3.2 show that the convergence rates of the noniterative estimators TOPUP and TIPUP can be improved by their iterative counterparts. Particularly, when the dimensions r_k for the factor process are fixed and the respective signal strength conditions are fulfilled, the proposed iTOPUP and iTIPUP just need one iteration to achieve the much sharper ideal rate $R^{(\mathrm{ideal})}$ in (3.10) and $R^{*(\mathrm{ideal})}$ in (3.17), compared with the rate (3.9) of TOPUP and (3.16) of TIPUP derived in Chen, Yang and Zhang (2022a), respectively. The improvement is achieved through replacing the much larger d_{-k} by r_{-k}, via orthogonal projection. When the factors are strong with $\delta_0 = \delta_1 = 0$ and the factor dimensions are fixed, the noniterative TOPUP-based estimators of Lam, Yao and Bathia (2011) for the vector factor model, Wang, Liu and Chen (2019) for the matrix factor and Chen, Yang and Zhang (2022a) for tensor factor models all have the same $O_{\mathbb{P}}(T^{-1/2})$ convergence rate for estimating the loading space. In comparison, the convergence rate $O_{\mathbb{P}}(T^{-1/2}d_{-k}^{-1/2})$ of both iterative estimators, iTOPUP and iTIPUP (when there is no severe signal cancellation, with bounded δ_2), is much sharper. Intuitively, when the signal is strong, the orthogonal projection operation helps to consolidate signals while potentially averaging out the noises, when the projection reduces the dimension of the mode-k unfolded matrix from $d_k \times d_{-k}$ for the tensor $\mathcal{X}_t$ to $d_k \times r_{-k}$ for the projected tensor $\mathcal{Z}_t$, resulting in the improvement by a factor of $d_{-k}^{-1/2}$ in the convergence rate.

When r_k are allowed to diverge, the iTOTUP and iTIPUP algorithms converge after at most $O(\log(d))$ iterations to achieve the ideal rate according to Theorems 3.1 and 3.2. The number of iterations needed can be as few as $O(\log(r))$ when the condition is right.

3.5.2. *Comparison between iTIPUP and iTOPUP.* The inner product operation in (2.7) for TIPUP-related procedures enjoys significant amount of noise cancellation compared to the outer product operation in (2.4) for TOPUP-related procedures. Compared with iTOPUP, the benefit of noise cancellation of the iTIPUP procedure is still visible through the reduction of r_{-k} in (3.10) to $\sqrt{r_{-k}}$ in (3.17) in the ideal rates. However, this post-iteration benefit is much less pronounced compared with the reduction of d_{-k} in (3.7) for TOPUP to $\sqrt{d_{-k}}$ in (3.16) for TIPUP in the noniterative rates. Meanwhile, the potential for signal cancellation in the TIPUP related schemes persists as λ_k^* and λ_k are unchanged between the initial and ideal

rates. We note that the signal strength can be viewed as λ_k and λ_k^* in Theorems 3.1 and 3.2, respectively, for TOPUP/iTOPUP and TIPUP/iTIPUP, and that severe signal cancellation can be expressed as $\lambda_k^* \ll \lambda_k$. When r_{-k} are allowed to diverge to infinity, the impact of signal cancellation is expressed in terms of δ_2 in Assumption 3: The iTOPUP has a faster rate than the iTIPUP when $\delta_2 > 3/2$, or $\{0 \le \delta \le 3/2, d_k \gtrsim r r_{-k}^{2-2\delta_2}, d_{-k} \gtrsim r r_{-k}^{3-2\delta_2}\}$, and slower rate when $d_k \lesssim r_k r_{-k}^{2-2\delta_2}$, in view of Corollary 3.3 and 3.4. In Corollaries 3.1 and 3.2, iTOPUP and iTIPUP have the same convergence rate because Corollary 3.2 assumes that signal cancellation does not change convergence rate.

Our results seem to suggest that the mixed TIPUP-iTOPUP procedure would strike a good balance between the benefit of noise cancellation (e.g., smaller T for consistency) and the potential danger of signal cancellation (e.g., $\lambda_k^* \ll \lambda_k$) for the following four reasons: (1) The benefit of noise cancellation is much larger in the initialization, in terms of d_{-k}, in view of the rates $R_k^{(0)}$ in (3.7) and $R^{*(0)}$ in (3.16). (2) The first part of condition (3.22) for TIPUP-iTOPUP is weaker than the first part of condition (3.19) for TIPUP-iTIPUP. (3) The signal strength λ_k of the stronger TOPUP form is retained in the rate $R^{(\text{ideal})}$ after iTOPUP iteration. (4) As we will prove in Section 3.6, the sample size requirement for the TIPUP initialization is optimal in the sense that it matches a computational lower bound under suitable conditions. Our simulation results support this recommendation, especially for relatively small r_{-k}. Of course, if the sample size qualitatively justifies the condition $C_1^{(\text{TOPUP})} R^{(0)} \le (1-\rho)/4$ in (3.12) and/or if a possible signal cancellation is a significant concern, the TOPUP initiation should be used.

3.5.3. *Comparison with HOOI.* The signal-to-noise ratio (SNR) condition, or equivalently the sample size requirement is mainly used to ensure that the initial estimator has sufficiently small estimation error. Thus, the performance of iterative procedures is measured by both the SNR requirement and the error rate achieved. Consider fixed h_0 in the fixed rank case with $K = 3$ and $d_{\max} \asymp d^{1/K}$. In the fixed signal model where $\mathcal{M}_t = \mathcal{M}$ is fixed and deterministic in (1.1), applying HOOI to the average of $\mathcal{X}_t$ would require SNR $\lambda(T^{1/2}/\sigma) \ge C_0 d^{1/4}$ to achieve the loss of the order $(\sigma/T^{1/2})d_k^{1/2}/\lambda$ according to Zhang and Xia (2018), where $\sigma/T^{1/2}$ is viewed as the noise level for HOOI as it is the standard deviation of each element of the average tensor. In terms of the autocross-products, taking the average over $\mathcal{X}_t$ roughly amounts to taking the average of all $T(T-1)/2$ lagged products between $\mathcal{X}_{t-h}$ and $\mathcal{X}_t$, $1 \le t-h < t \le T$. However, in the tensor factor model (1.1) where the signal part is random and serial correlated, the average is taken only over $T-h$ lagged products for each h. Thus, while the rate of the average of the signal-by-noise cross-products in the factor model is heuristically expected to match that of HOOI at noise level $\sigma/T^{1/2}$, the rate of the average of the noise-by-noise cross-products in the factor model is expected to only match that of HOOI with noise level $\sigma/T^{1/4}$. In Corollary 3.2, the contribution of the noise-by-noise cross-products dominates the initial estimation error as the SNR requirement $\lambda(T^{1/4}/\sigma) \ge C_0 d^{1/4}$ in (3.25) matches that of HOOI with noise level $\sigma/T^{1/4}$; at the same time, the contribution of the signal-by-noise cross-products dominates the estimation error after iteration as the rate $(\sigma/T^{1/2})d_k^{1/2}/\lambda$ in (3.26) matches that of HOOI with noise level $\sigma/T^{1/2}$. Thus, if there is no severe signal cancellation, the signal-to-noise ratio requirement and convergence rate for iTIPUP and TIPIP-iTOPUP in the factor model are both comparable with those of HOOI in the simpler fixed signal setting, but the rate match is achieved in very different and subtle ways. We prove that this insight is intrinsic as the rates in (3.25) and (3.26) are both optimal according to the computational and statistical lower bounds in the following subsection.

3.6. *Computational and statistical lower bounds*. In this subsection, we focus on the typical factor model setting that the condition numbers of $A_k^\top A_k$ are bounded. We shall prove that under the computational hardness assumption, the signal-to-noise ratio condition (3.25) imposed on iTIPUP (also TIPUP-iTOPUP) in Corollary 3.2 is unavoidable for computationally feasible estimators to be consistent. To be specific, we show that, if the signal-to-noise ratio condition is violated, then any computationally efficient and consistent estimator of the loading spaces leads to a computationally efficient and statistically consistent test for the hypergraphic planted clique detection problem in a regime where it is believed to be computationally intractable. In addition, we establish a statistical lower bound on the minimax risk of the estimators.

Hypergraphic planted clique. An m-hypergraph $G = (V(G), E(G))$ is a natural extension of regular graph, where $V(G) = [N]$ and each hyperedge is represented by an unordered group of m different vertices $i_j \in V(G)$ $(j = 1, \ldots, m)$, denoted as $e = (i_1, \ldots, i_m) \in E(G)$. Given a m-hypergraph its adjacency tensor $\mathcal{A} \in \{0, 1\}^{N \times N \times \cdots \times N}$ is defined as

$$\mathcal{A}_{i_1,\ldots,i_m} = \begin{cases} 1 & \text{if } e = (i_1, \ldots, i_m) \in E(G); \\ 0 & \text{otherwise.} \end{cases}$$

We denote by $\mathcal{G}_m(N, 1/2)$ the Erdős–Rényi m-hypergraph on N vertices where each hyperedge e is drawn independently with probability $1/2$, by $\mathcal{C} = \mathcal{C}(N, \kappa)$ a random clique of size κ where the κ members are uniformly sampled from $[N]$ and $E(\mathcal{C})$ is composed of all $e = (i_1, \ldots, i_m)$ with $i_j \in \mathcal{C}$, and by $\mathcal{G}_m(N, 1/2, \kappa)$ the random graph generated by first sampling independently $\mathcal{G}_m(N, 1/2)$ and $\mathcal{C} = \mathcal{C}(N, \kappa)$ and then adding all the edges in $E(\mathcal{C})$ to the set of edges in $\mathcal{G}_m(N, 1/2)$. The Hypergraphic Planted Clique (HPC) detection problem of parameter (N, κ, m) refers to testing the following hypotheses:

$$H_0^G : \mathcal{A} \sim \mathcal{G}_m(N, 1/2) \quad \text{vs.} \quad H_1^G : \mathcal{A} \sim \mathcal{G}_m(N, 1/2, \kappa). \tag{3.32}$$

If $m = 2$, the above HPC detection becomes the traditional planted clique (PC) detection problem. When $\kappa \geq c\sqrt{N}$, many computationally efficient algorithms have been developed for PC detection; see, Alon, Krivelevich and Sudakov (1998), Feige and Krauthgamer (2000), Feige and Ron (2010), Ames and Vavasis (2011), Dekel, Gurel-Gurevich and Peres (2014), Deshpande and Montanari (2015), Feldman et al. (2017), among others. However, it has been widely conjectured that when $\kappa = o(\sqrt{N})$, the PC detection problem cannot be solved in randomized polynomial time, which is referred to as the hardness conjecture. Computational lower bounds in several statistical problems have been established by assuming the hardness conjecture of PC detection, including sparse PCA (Berthet and Rigollet (2013a,b), Wang, Berthet and Samworth (2016)), sparse CCA (Gao, Ma and Zhou (2017)), submatrix detection (Ma and Wu (2015), Cai, Liang and Rakhlin (2017)), community detection (Hajek, Wu and Xu (2015)), etc.

Recently, motivated by tensor data analysis, hardness conjecture for HPC detection problem has been proposed; see, for example, Zhang and Xia (2018), Brennan and Bresler (2020), Luo and Zhang (2022, 2020), Pananjady and Samworth (2022). Similar to the PC detection, they hypothesized that when $\kappa = O(N^{1/2-\delta})$ with $\delta > 0$, the HPC detection problem (3.32) cannot be solved by any randomized polynomial-time algorithm. Formally, the conjectured hardness of the HPC detection problem can be stated as follows.

HYPOTHESIS I (HPC detection). *Consider the HPC detection problem* (3.32) *and suppose* $m \geq 2$ *is a fixed integer. If*

$$\limsup_{N \to \infty} \frac{\log \kappa}{\log N} \leq \frac{1}{2} - \delta, \quad \textit{for any } \delta > 0, \tag{3.33}$$

for any sequence of polynomial-time tests $\{\psi\}_N : \mathcal{A} \to \{0,1\}$,

$$\limsup_{N\to\infty}\big(\mathbb{P}_{H_0^G}(\psi(\mathcal{A})=1)+\mathbb{P}_{H_1^G}(\psi(\mathcal{A})=0)\big)>1/2.$$

Evidence supporting this hypothesis has been provided in Zhang and Xia (2018), Luo and Zhang (2022). This version of the hypothesis is similar to the one in Berthet and Rigollet (2013a), Ma and Wu (2015), Gao, Ma and Zhou (2017) for the PC detection problem.

For simplicity, we especially consider the factor model (1.2) with each individual series of $\mathcal{F}_t$ being mean 0 and independent,

$$\mathcal{X}_t = \lambda \mathcal{F}_t \times_1 U_1 \times_2 \ldots \times_K U_K + \mathcal{E}_t, \tag{3.34}$$

where $U_k \in \mathbb{R}^{d_k\times r_k}$, $U_k^\top U_k = I$ for $1\le k\le K$ and $0<c_1\le \sigma_{\min}(\mathbb{E}\vec{1}(\mathcal{F}_t)\vec{1}^\top(\mathcal{F}_t))\le \sigma_{\max}(\mathbb{E}\vec{1}(\mathcal{F}_t)\vec{1}^\top(\mathcal{F}_t))\le c_2<\infty$. The probability space we consider in this section is

$$\begin{aligned}
&\mathscr{P}(T,d_1,\ldots,d_K,\lambda)\\
&\quad=\Bigg\{\mathcal{X}_1,\ldots,\mathcal{X}_T : \mathcal{X}_t \text{ has form (3.34) with independent series } \mathcal{F}_{t,i_1,\ldots,i_K},\\
&\qquad \frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}\mathcal{F}_{t,i_1,\ldots,i_K}\mathcal{F}_{t-1,i_1,\ldots,i_K}=c_0>0,\\
&\qquad \text{and } \{\mathcal{F}_t\}_{t=1}^T \text{ independent of } \{\mathcal{E}_t\}_{t=1}^T,\\
&\qquad \mathcal{E}_{t,j_1,\ldots,j_K}\overset{\text{i.i.d.}}{\sim} N(0,\sigma^2),\\
&\qquad \text{for all } 1\le t\le T, 1\le i_k\le r_k, 1\le j_k\le d_k, 1\le k\le K\Bigg\}.
\end{aligned} \tag{3.35}$$

The computational lower bound over $\mathscr{P}(T,d_1,\ldots,d_K,\lambda)$ is then presented as below. In the case of $K=1$, the problem reduces to PCA of the spike covariance matrix. For general K, the autocovariance tensor is of order $2K$.

THEOREM 3.4. *Suppose that Hypothesis* I *holds for some* $0<\delta<1/2$ *and* $d^{1/K}\asymp d_k\ge T$ *and* r_k *is fixed for all* $1\le k\le K$. *If, for some* $\vartheta>0$,

$$\liminf_{T\to\infty}\frac{\sigma^2 d^{1/2-\vartheta}}{T^{1/2}\lambda^2}>0, \tag{3.36}$$

then for any randomized polynomial-time estimators $\widehat{U}_k=\widehat{U}_k(\mathcal{X}_1,\ldots,\mathcal{X}_T)$, $1\le k\le K$,

$$\liminf_{T\to\infty}\sup_{\mathcal{X}_1,\ldots,\mathcal{X}_T\in\mathscr{P}(T,d_1,\ldots,d_K,\lambda)}\mathbb{P}\Big(\min_{1\le k\le K}\|\widehat{P}_k-P_k\|_{\mathrm{S}}^2>\frac{1}{3}\Big)>\frac{1}{4}, \tag{3.37}$$

where $\widehat{P}_k=\widehat{U}_k\widehat{U}_k^\top$ *and* $P_k=U_kU_k^\top$.

Comparing (3.36) with (3.25), we see that the signal-to-noise ratio condition (3.25) cannot be improved upon by a factor of d^ϑ with polynomial time complexity for any $\vartheta>0$. The condition $d_k\ge T$ is a technical requirement to use the theoretical tools in Ma and Wu (2015) and Brennan and Bresler (2020) for the reduction from HPC.

REMARK 3.11. Theorem 3.4 illustrates the computational hardness for factor loading spaces estimation under the typical factor model setting that the condition numbers of $A_k^\top A_k$ are bounded and ranks r_k are fixed, and suggests the use of TIPUP initialization with proper fixed h_0 as it attains the computational lower bound under the typical factor model setting.

REMARK 3.12. In the general r_k case, the optimal signal-to-noise ratio requirement falls between $\lambda^2/\sigma^2 \gtrsim \max_{1\le k\le K}\sqrt{d}/(\sqrt{T}r_{-k})$ (by Theorem 3.4) and $\lambda^2/\sigma^2 \gtrsim \max_{1\le k\le K} d_k/(Tr_{-k})$ (by Theorem 3.5 below). It seems possible to unfold tensor into matrix and use the results of Ma and Wu (2015) to narrow the gap; however, a complete solution to this challenging problem is beyond the scope of our paper.

Next, we establish the statistical lower bound for the tensor factor model problem. Again, we consider the probability space (3.35).

THEOREM 3.5. *Suppose $\lambda > 0$ and $d_k \to \infty$ as $T \to \infty$ for all $1 \le k \le K$. Then there exists a universal constant $c > 0$ such that for T sufficiently large,*

$$\inf_{\widehat{U}_k}\sup_{\mathcal{X}_1,\ldots,\mathcal{X}_T\in\mathscr{P}(T,d_1,\ldots,d_K,\lambda)}\mathbb{E}\|\widehat{P}_k - P_k\|_{\mathrm{S}} \ge c\min\bigl(1,(\sigma^2+\sigma\lambda)\sqrt{d_k}/(\lambda^2\sqrt{Tr_{-k}})\bigr) \tag{3.38}$$

for all $1 \le k \le K$, where $\widehat{P}_k = \widehat{U}_k\widehat{U}_k^\top$ and $P_k = U_kU_k^\top$.

REMARK 3.13. The statistical lower bound for high-dimensional tensor factor models is provided in Theorem 3.5. This bound directly matches the upper bounds in Corollary 3.2 and also matches the bounds in Corollary 3.1 when $d_k \gtrsim rr_{-k}^2$ and $\lambda^2/\sigma^2 \gtrsim d_kr_{-k}^5/T + d_k^{1/2}r_{-k}^{3/2}/T^{1/2}$. These results demonstrate that the rates obtained by our proposed iterative procedures are minimax-optimal. Moreover, Theorem 3.5 reveals a different effect of the ranks r_k ($k = 1,\ldots,K$) compared to tensor Tucker decomposition (Zhang and Xia (2018)), further confirming the distinct nature of tensor factor models from low-rank matrix/tensor problems.

4. A matrix perturbation bound. In Lemma 4.1 below, we provide an improvement of the matrix perturbation bound of Wedin (1972). The lemma, proved in Appendix G in the Supplementary Material and used to prove Proposition 3.1, is of independent interest due to wide applications of the Wedin (1972) bound.

LEMMA 4.1. *Let $r \le d_1 \wedge d_2$, M be a $d_1 \times d_2$ matrix, U and V be respectively the left and right singular matrices associated with the r largest singular values of M, $U_\perp$ and $V_\perp$ be the orthonormal complements of U and V, and λ_r be the rth largest singular value of M. Let $\widehat{M} = M + \Delta$ be a noisy version of M, $\{\widehat{U},\widehat{V},\widehat{U}_\perp,\widehat{V}_\perp\}$ be the counterpart of $\{U,V,V_\perp,V_\perp\}$ and $\widehat{\lambda}_{r+1}$ be the $(r+1)$th largest singular value of $\widehat{M}$. Let $\|\cdot\|$ be a matrix norm satisfying $\|ABC\| \le \|A\|_{\mathrm{S}}\|C\|_{\mathrm{S}}\|B\|$, $\epsilon_1 = \|U^\top\Delta\widehat{V}_\perp\|$ and $\epsilon_2 = \|\widehat{U}_\perp^\top\Delta V\|$. Then*

$$\|U_\perp^\top\widehat{U}\| \le \frac{\widehat{\lambda}_{r+1}\epsilon_1 + \lambda_r\epsilon_2}{\lambda_r^2 - \widehat{\lambda}_{r+1}^2} \le \frac{\epsilon_1\vee\epsilon_2}{\lambda_r - \widehat{\lambda}_{r+1}}. \tag{4.1}$$

In particular, for the spectral norm $\|\cdot\| = \|\cdot\|_{\mathrm{S}}$, $\mathrm{error}_1 = \|\Delta\|_S/\lambda_r$ and $\mathrm{error}_2 = \epsilon_2/\lambda_r$,

$$\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathrm{S}} \le \frac{\mathrm{error}_1^2 + \mathrm{error}_2}{1 - \mathrm{error}_1^2}. \tag{4.2}$$

The sharper perturbation bound in the middle of (4.1) improves the commonly used version of the Wedin (1972) bound on the right-hand side, compared with Theorem 1 of Cai and Zhang (2018) and Lemma 1 of Chen, Yang and Zhang (2022b). As Cai and Zhang (2018) pointed out, such variations of the Wedin (1972) bound provide sharper convergence rate when $\mathrm{error}_2 \le \mathrm{error}_1$ in (4.2), typically in the case of $d_1 \ll d_2$, as in Proposition 3.1.

5. Summary. In this paper, we propose new estimation procedures for the tensor factor model via iterative projection, and focus on two procedures: iTOPUP and iTIPUP. Theoretical analysis shows the asymptotic properties of the estimators. Simulation study presented in the Supplementary Material illustrates the finite sample properties of the estimators. While theoretical results are obtained under very general conditions, concrete specific cases are considered. In particular, under the typical factor model setting where the condition numbers of $A_k^\top A_k$ are bounded and the ranks r_k are fixed, the proposed iterative procedures, iTOPUP method and iTIPUP method (with no severe signal cancellation) lead to a convergence rate $O_{\mathbb{P}}((Td_{-k})^{-1/2})$ under strong factors settings due to information pooling of the orthogonal projection of the other d_{-k} dimensions. This rate is much sharper than the existing rate $O_{\mathbb{P}}(T^{-1/2})$ in the recent literature for noniterative estimators for vector, matrix and tensor factor models. It implies that the accuracy can be improved by increasing the dimensions, and consistent estimation of the loading spaces can be achieved even with a fixed finite sample size T. This is in sharp contrast to the folklore based on the existing literature that only the sample size T helps the estimation of the loading matrices in factor models. The proposed iterative estimation methods not only preserve the tensor structure, but also result in sharper convergence rate in the estimation of factor loading space.

The iterative procedure requires two operators, one for initialization and one for iteration. Under certain conditions of the signal-to-noise ratio (or the sample size requirement), we only need the initial estimator to have sufficiently small estimation errors but not the consistency of the initial estimator. Often, one iteration is sufficient. In more complicated general cases, at most $O(\log(d))$ iterations are needed to achieve the ideal rate of convergence. Based on the theoretical results and empirical evidence, we suggest to use iTOPUP for iteration when the ranks r_k are small. In terms of initiation, the computational lower bound shows that the signal-to-noise ratio condition derived from TIPUP initialization is unavoidable for any computationally feasible estimation procedure to achieve consistency, while that from TOPUP initialization is not optimal. Based on this result, we suggest the use of TIPUP initialization. Of course, this should be done with precaution against potential signal cancellation, for example, by using a slightly large h_0 as our empirical results show. By examination of the patterns of estimated singular values under different lag values h_0, using iTOPUP and iTIPUP, it is possible to detect signal cancellation, which has significant impact on iTIPUP estimators.

The proposed iterative procedure is similar to HOOI algorithms in spirit, but the detailed operations and the theoretical challenges are significantly different.

Acknowledgments. We would like to thank the Editor, the Associate Editor and the anonymous referees for their detailed reviews, which helped to improve the paper substantially.

Funding. Yuefeng Han's research is supported in part by National Science Foundation Grant IIS-1741390.

Rong Chen's research is supported in part by National Science Foundation Grants DMS-1737857, IIS-1741390, CCF-1934924, DMS-2027855 and DMS-2319260.

Dan Yang's research is supported in part by NSF Grant IIS-1741390, Hong Kong Grant GRF 17301620, Hong Kong Grant CRF C7162-20GF and Shenzhen Grant SZRI2023-TBRF-03.

Cun-Hui Zhang's research is supported in part by NSF Grants DMS-1721495, IIS-1741390, CCF-1934924, DMS-2052949 and DMS-2210850.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Tensor factor model estimation by iterative projection" (DOI: 10.1214/24-AOS2412SUPP; .pdf). In the supplementary material, we provide simulation studies, the proofs of main results in the paper and some lemmas that are useful in proofs of the paper.

REFERENCES

AHN, S. C. and HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227. MR3064065 https://doi.org/10.3982/ECTA8968

ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. *Random Structures Algorithms* **13** 457–466.

ALTER, O. and GOLUB, G. H. (2005). Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations. *Proc. Natl. Acad. Sci. USA* **102** 17559–17564.

AMES, B. P. W. and VAVASIS, S. A. (2011). Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.* **129** 69–89. MR2831403 https://doi.org/10.1007/s10107-011-0459-x

ANANDKUMAR, A., GE, R., HSU, D. and KAKADE, S. M. (2014). A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.* **15** 2239–2312. MR3231594

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857 https://doi.org/10.1111/1468-0262.00392

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273

BAI, J. and NG, S. (2007). Determining the number of primitive shocks in factor models. *J. Bus. Econom. Statist.* **25** 52–60. MR2338870 https://doi.org/10.1198/073500106000000413

BERTHET, Q. and RIGOLLET, P. (2013a). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066. PMLR.

BERTHET, Q. and RIGOLLET, P. (2013b). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849 https://doi.org/10.1214/13-AOS1127

BI, X., QU, A. and SHEN, X. (2018). Multilayer tensor factorization with applications to recommender systems. *Ann. Statist.* **46** 3308–3333. MR3852653 https://doi.org/10.1214/17-AOS1659

BRENNAN, M. and BRESLER, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory* 648–847. PMLR.

CAI, T. T., LIANG, T. and RAKHLIN, A. (2017). Computational and statistical boundaries for submatrix localization in a large noisy matrix. *Ann. Statist.* **45** 1403–1430. MR3670183 https://doi.org/10.1214/16-AOS1488

CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89. MR3766946 https://doi.org/10.1214/17-AOS1541

CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050 https://doi.org/10.2307/1912275

CHANG, J., HE, J., YANG, L. and YAO, Q. (2023). Modelling matrix time series via a tensor CP-decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 127–148. MR4726961 https://doi.org/10.1093/jrsssb/qkac011

CHEN, E. Y. and CHEN, R. (2022). Modeling dynamic transport network with matrix factor models: An application to international trade flow. *J. Data Sci.* **21** 490–507.

CHEN, E. Y. and FAN, J. (2023). Statistical inference for high-dimensional matrix-variate factor models. *J. Amer. Statist. Assoc.* **118** 1038–1055. MR4595475 https://doi.org/10.1080/01621459.2021.1970569

CHEN, E. Y., TSAY, R. S. and CHEN, R. (2020). Constrained factor models for high-dimensional matrix-variate time series. *J. Amer. Statist. Assoc.* **115** 775–793. MR4107679 https://doi.org/10.1080/01621459.2019.1584899

CHEN, E. Y., XIA, D., CAI, C. and FAN, J. (2024). Semi-parametric tensor factor analysis by iteratively projected singular value decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 793–823. MR4792086 https://doi.org/10.1093/jrsssb/qkae001

CHEN, R., YANG, D. and ZHANG, C.-H. (2022a). Factor models for high-dimensional tensor time series. *J. Amer. Statist. Assoc.* **117** 94–116. MR4399070 https://doi.org/10.1080/01621459.2021.1912757

CHEN, R., YANG, D. and ZHANG, C.-H. (2022b). Rejoinder. *J. Amer. Statist. Assoc.* **117** 128–132.

DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21** 1324–1342. MR1780276 https://doi.org/10.1137/S0895479898346995

DEKEL, Y., GUREL-GUREVICH, O. and PERES, Y. (2014). Finding hidden cliques in linear time with high probability. *Combin. Probab. Comput.* **23** 29–49. MR3197965 https://doi.org/10.1017/S096354831300045X

DESHPANDE, Y. and MONTANARI, A. (2015). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Found. Comput. Math.* **15** 1069–1128. MR3371378 https://doi.org/10.1007/s10208-014-9215-y

FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410 https://doi.org/10.1214/11-AOS944

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016

FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254. MR3449767 https://doi.org/10.1214/15-AOS1364

FAN, J., LIU, H. and WANG, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.* **46** 1383–1414. MR3819104 https://doi.org/10.1214/17-AOS1588

FEIGE, U. and KRAUTHGAMER, R. (2000). Finding and certifying a large hidden clique in a semirandom graph. *Random Structures Algorithms* **16** 195–208. MR1742351 https://doi.org/10.1002/(SICI)1098-2418(200003)16:2<195::AID-RSA5>3.3.CO;2-1

FEIGE, U. and RON, D. (2010). Finding hidden cliques in linear time. In 21*st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms* (*AofA*'10). *Discrete Math. Theor. Comput. Sci. Proc., AM* 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy. MR2735341

FELDMAN, V., GRIGORESCU, E., REYZIN, L., VEMPALA, S. S. and XIAO, Y. (2017). Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM* **64** 8. MR3664576 https://doi.org/10.1145/3046674

FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Rev. Econ. Stat.* **82** 540–554.

FOSTER, G. (1996). Time series analysis by projection. II. Tensor methods for time series analysis. *Astron. J.* **111** 555.

GAO, C., MA, Z. and ZHOU, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45** 2074–2101. MR3718162 https://doi.org/10.1214/16-AOS1519

HAJEK, B., WU, Y. and XU, J. (2015). Computational lower bounds for community detection on random graphs. In *Conference on Learning Theory* 899–928. PMLR.

HALLIN, M. and LIŠKA, R. (2007). Determining the number of factors in the general dynamic factor model. *J. Amer. Statist. Assoc.* **102** 603–617. MR2325115 https://doi.org/10.1198/016214506000001275

HAN, Y., CHEN, R., YANG, D. and ZHANG, C.-H (2024). Supplement to "Tensor Factor Model Estimation by Iterative Projection." https://doi.org/10.1214/24-AOS2412SUPP

HAN, Y. and ZHANG, C.-H. (2023). Tensor principal component analysis in high dimensional CP models. *IEEE Trans. Inf. Theory* **69** 1147–1167. MR4564648 https://doi.org/10.1109/tit.2022.3203972

HAN, Y., ZHANG, C.-H. and CHEN, R. (2021). CP factor model for dynamic tensors. ArXiv Preprint. Available at arXiv:2110.15517.

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663 https://doi.org/10.1214/12-AOS970

LAM, C., YAO, Q. and BATHIA, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika* **98** 901–918. MR2860332 https://doi.org/10.1093/biomet/asr048

LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2012). Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 208–220.

LIU, Y., SHANG, F., FAN, W., CHENG, J. and CHENG, H. (2014). Generalized higher-order orthogonal iteration for tensor decomposition and completion. In *Advances in Neural Information Processing Systems* 1763–1771.

LUO, Y. and ZHANG, A. R. (2020). Open problem: Average-case hardness of hypergraphic planted clique detection. In *Conference on Learning Theory* 3852–3856. PMLR.

LUO, Y. and ZHANG, A. R. (2022). Tensor clustering with planted structures: Statistical optimality and computational limits. *Ann. Statist.* **50** 584–613. MR4382029 https://doi.org/10.1214/21-aos2123

MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116. MR3346698 https://doi.org/10.1214/14-AOS1300

OMBERG, L., GOLUB, G. H. and ALTER, O. (2007). A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc. Natl. Acad. Sci. USA* **104** 18371–18376.

OUYANG, J. and YUAN, M. (2022). Comments on "Factor models for high-dimensional tensor time series" [4399070]. *J. Amer. Statist. Assoc.* **117** 124–127. MR4399073 https://doi.org/10.1080/01621459.2022.2028630

PAN, J. and YAO, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95** 365–379. MR2521589 https://doi.org/10.1093/biomet/asn009

PANANJADY, A. and SAMWORTH, R. J. (2022). Isotonic regression with unknown permutations: Statistics, computation and adaptation. *Ann. Statist.* **50** 324–350. MR4382019 https://doi.org/10.1214/21-aos2107

PEÑA, D. and BOX, G. E. P. (1987). Identifying a simplifying structure in time series. *J. Amer. Statist. Assoc.* **82** 836–843. MR0909990

ROGERS, M., LI, L. and RUSSELL, S. J. (2013). Multilinear dynamical systems for tensor time series. *Adv. Neural Inf. Process. Syst.* **26** 2634–2642.

SHEEHAN, B. N. and SAAD, Y. (2007). Higher order orthogonal iteration of tensors (HOOI) and its relation to PCA and GLRAM. In *Proceedings of the* 2007 *SIAM International Conference on Data Mining* 355–365. SIAM, Philadelphia.

STOCK, J. H. and WATSON, M. W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97** 1167–1179. MR1951271 https://doi.org/10.1198/016214502388618960

SUN, W. W. and LI, L. (2017). STORE: Sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18** 135. MR3763769

TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. MR0205395 https://doi.org/10.1007/BF02289464

WANG, D., LIU, X. and CHEN, R. (2019). Factor models for matrix-valued high-dimensional time series. *J. Econometrics* **208** 231–248. MR3906969 https://doi.org/10.1016/j.jeconom.2018.09.013

WANG, D., ZHENG, Y. and LI, G. (2024). High-dimensional low-rank tensor autoregressive time series modeling. *J. Econometrics* **238** 105544. MR4663431 https://doi.org/10.1016/j.jeconom.2023.105544

WANG, T., BERTHET, Q. and SAMWORTH, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.* **44** 1896–1930. MR3546438 https://doi.org/10.1214/15-AOS1369

WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968 https://doi.org/10.1007/bf01932678

ZHANG, A. and XIA, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Trans. Inf. Theory* **64** 7311–7338. MR3876445 https://doi.org/10.1109/TIT.2018.2841377

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640 https://doi.org/10.1080/01621459.2013.776499