Optimal Distributed Tensor PCA with Statistical Guarantees for Heterogeneous and Decentralized Data

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 6, 2024

Abstract

In this paper, we propose new methodologies for distributed tensor Principal Component Analysis (PCA) in heterogeneous data environments. These environments often arise when tensor datasets are stored across multiple machines, each with unique statistical properties, presenting significant challenges for aggregation and analysis. We extend classical Tucker decomposition to a distributed setting, providing sharp statistical convergence guarantees. Our approach handles both homogeneous and heterogeneous tensor data by developing estimators that account for common and individual components, achieving minimax optimal rates in various data settings. Extensive numerical simulations and real data analysis demonstrate the efficacy of our methods, which outperform traditional single-location PCA and pooled estimators, especially in heterogeneous environments. Key applications include healthcare data analysis and recommendation systems, where privacy and communication costs are critical factors.

Keywords: Distributed Tensor PCA, Tucker Decomposition, Heterogeneous Data, Minimax Optimality, Statistical Guarantees.

1 Introduction

The rapid growth of large-scale tensor datasets across various domains, such as healthcare systems, recommendation systems, and neuroimaging, necessitates efficient methodologies capable of handling decentralized data storage and heterogeneous environments. Traditional tensor Principal Component Analysis (PCA) techniques, such as Tucker decomposition, have been extensively applied for dimension reduction and unsupervised learning in centralized settings. However, with the increasing need for privacy preservation, communication cost reduction, and concerns over data ownership, new approaches are required for distributed environments where data pooling is not feasible.

In this paper, we address these challenges by proposing novel algorithms for distributed tensor PCA that extend classical Tucker decomposition to both homogeneous and heterogeneous data settings. Our methods are designed to estimate both the common and unique components of tensor datasets distributed across multiple locations, leveraging efficient communication protocols that minimize data transfer. This is particularly crucial in settings where tensors have varying underlying structures, such as medical records dispersed across hospitals or customer data spread across global IT firms. Furthermore, our approach achieves sharp statistical convergence guarantees, ensuring both accuracy and efficiency in decentralized environments.

Brief Contributions Our key contributions include:

- We propose a novel distributed tensor PCA framework that efficiently handles heterogeneous environments by accommodating both common and individual components of tensors across machines.
- We establish sharp statistical error bounds, demonstrating that our methods achieve minimax optimal rates in various data settings, including highly heterogeneous scenarios.
- We develop a transfer learning approach to enhance tensor estimation at a target location by leveraging data from other distributed sites, optimizing the balance between noise levels and data relevance.
- Extensive numerical simulations and real-world experiments validate the effectiveness of our methods, showing improved estimation accuracy and communication efficiency compared to traditional approaches.

1.1 Motivations and Contributions

In recent years, the prevalence of large-scale tensor datasets across diverse fields has garnered significant attention. Tucker low-rank tensor Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) have become indispensable for unsupervised learning and dimension reduction, proving crucial in deriving valuable insights across various applications. These techniques are extensively utilized in recommendation systems [KABO10], natural image and video processing [GdSK$^+$21], neuroimaging analysis [LXZL18], and health care systems [RYZ$^+$23], among others.

Despite their wide applicability, the geographical dispersion of large-scale tensor datasets presents substantial challenges in data aggregation and analysis due to high communication costs, privacy concerns, and data security and ownership issues. For example, IT corporations face practical limitations in centralizing globally gathered customer data for recommendation systems due to communication budgets and network bandwidth constraints. Similarly, health records spread across multiple hospitals or jurisdictions pose significant privacy and ownership challenges for centralized processing.

Addressing these challenges, our paper makes several critical contributions to the field of tensor decomposition, particularly focusing on its application in distributed and heterogeneous environments. While recent studies have begun exploring distributed Tucker decomposition, they primarily concentrate on computational aspects such as parallelization and memory usage, focusing on environments where tensors share identical principal spaces [SSK16, CCJ$^+$18, JK20]. These studies often overlook the importance of providing statistical guarantees and also fail to address the challenges posed by heterogeneous tensors, especially prevalent in fields like medical care, where tensor data from various locations differ in their underlying structures. Our research fills this gap by not only providing statistical convergence and inference theories for distributed Tucker decomposition but also extending its applicability beyond homogeneous data settings.

In this paper, we address the challenge of tensor PCA within a distributed framework, where tensors are stored across different machines without the feasibility of pooling them together. Our comprehensive analysis delineates three distinct scenarios: In the homogeneous scenario in Section 2, we develop a method that involves computing local estimators for each tensor's subspace, then transmitting these estimators to a central processor, and finally aggregating them to produce a global estimate, reducing communication by only transmitting essential subspace information.

In the second scenario with heterogeneity in Section 3, tensors observed on different machines are allowed to be generated from different underlying models that share some common principal

components. Our focus is to improve estimation across all machines handling heterogeneous data. We formally define the partition of common and individual components for each tensor and develop a generalized distributed Tensor PCA method to efficiently estimate these shared components and also identify and extract unique components specific to each tensor. This approach allows us to accommodate the distinct characteristics of each dataset within the collective framework.

The third scenario with heterogeneity in Section 4 focuses on another setting of heterogeneous distributed learning, aiming at enhancing the estimation accuracy of a tensor at a specific location by intelligently integrating abundant data from other locations. We develop a novel transfer learning algorithm based on a weighted averaging scheme that involves calibrating the influence of data based on their noise levels and relevance, thereby improving the precision of the target site's tensor estimation.

We have established comprehensive statistical error bounds for our proposed estimators across different scenarios. In Section 2.3, our analysis involves a detailed decomposition and quantification of bias and variance terms. We demonstrate that our estimator achieves the optimal minimax rate when the signal-to-noise ratio (SNR) exceeds a certain threshold, thus equating the performance of the pooled estimator that aggregates tensors directly (further details in Section 2.1). Essentially, our distributed algorithm attains the highest possible estimation accuracy achievable in a non-distributed setting.

In the heterogeneous scenario detailed in Section 3.2, we show that our estimator for common components matches the performance rate of the homogeneous scenario, while the estimators for individual components maintain rates consistent with individual tensor PCA. This underscores the effectiveness of our method in simultaneously learning shared and unique tensor structures. Importantly, even when pooling of tensors is possible, creating an efficient estimator through direct aggregation is challenging due to data heterogeneity—a fact corroborated by our numerical studies where our distributed methods surpass the pooled estimator in various heterogeneous configurations.

Furthermore, we provide theoretical guarantees for our estimator designed to utilize knowledge from multiple distributed sites, optimizing the weight allocations in our algorithm to enhance estimation accuracy beyond what is achievable with the target dataset alone. In addition, we derive the asymptotic distribution of the distance between our proposed estimator and the true values in Section B of the supplementary material, facilitating statistical inference and enabling the construction of confidence regions for the singular subspaces.

Our extensive numerical evaluations, detailed in Section 5, assess the empirical performance of our methods. The results demonstrate that our approaches not only significantly outperform the "single" method, which applies PCA on individual tensors without aggregation, but also exceed the performance of the pooled method in both simulated heterogeneous settings and real data analyses. These findings highlight the superiority of our distributed approaches, which enhance estimation accuracy by effectively aggregating common information across diverse tensors amid data heterogeneity.

Our work, while related to distributed matrix PCA, tackles the more complex issue of distributed tensor PCA, which is inherently more challenging both practically and technically. Unlike matrices, tensors often involve three or more modes, increasing their dimensionality and complicating their analysis. Whereas matrices might be addressed through convex methods like nuclear norm penalization, tensors require nonconvex formulations due to their more complex structure, involving multiple low-rank components and a core tensor. This introduces significant challenges in algorithm design and the establishment of theoretical guarantees, including statistical convergence

and inference, which are critical for distributed tensor PCA. The contributions of our work are summarized as follows.

- *Modeling*: We introduce a new model to represent the distributed and heterogeneous environments encountered in tensor PCA. This model addresses a gap in the current literature, which has largely overlooked the complexities of real-world distributed data analysis where tensors are heterogeneous on different machines or servers.

- *Methodological Advances*: We propose novel distributed methods for tensor PCA that function effectively under both homogeneous and heterogeneous settings. These methods are designed to aggregate common components shared across different tensors, thereby enhancing estimation accuracy. Additionally, we expand our methodology to address scenarios akin to transfer learning, making considerable strides in distributed tensor analysis. To the best of our knowledge, we are the first to develop such methods for distributed tensor analysis, which, as discussed above, are highly different from the existing methods for matrix PCA.

- *Theoretical Contributions*: We establish statistical guarantees for our distributed methods, demonstrating that they achieve a sharp statistical error rate that aligns with the minimax optimal rate possible in a non-distributed setting. Additionally, we calculate the asymptotic distribution of the distance between our estimator and the true values, as detailed in Section B of the supplementary material. This calculation aids in statistical inference and supports the construction of confidence regions for the singular subspaces. These contributions represent a theoretical advancement in utilizing aggregated common components, a finding that has not been previously documented in tensor analysis literature and enhances the existing computational approaches to distributed Tucker decomposition.

1.2 Related works

This paper is situated at the intersection of two bodies of literature: tensor decomposition and distributed estimation and inference. The literature on both areas is broad and vast. Here we only review the closely related studies, namely tensor decomposition and distributed learning that provides statistical guarantees. The readers are referred to [BTY$^+$21] and [SHL21] for recent surveys of statistical tensor learning.

Tensor Decomposition. Motivated by modern scientific research, tensor decompositions have been actively studied in machine learning, electrical engineering, and statistics. Despite the well-established theory for low-rank decomposition of matrices, tensors present unique challenges. There are several notions of low-rankness in tensors [KB09], including the most popular CANDECOMP/PARAFAC (CP) low-rankness and multilinear/Tucker low-rankness. [RM14], [HSS15], and [PWB20] considered a rank-1 spiked statistical tensor PCA model, a special case of CP decomposition, and proposed various methods, including tensor unfolding and sum of square optimization. [ZX18] introduced a general framework of tensor SVD based on Tucker decomposition and established statistical and computational limits, followed by [XZZ22] who further studied inference for tensor PCA. [ZH19] considered sparse tensor SVD where the loading matrices are assumed to be sparse. [WL20] studied the decomposition of a higher-order tensor with binary entries. [CXCF24] further introduced a general framework of semi-parametric tensor factor models based on tensor PCA with auxiliary covariate information.

Distributed Estimation and Inference. To handle the challenges posed by massive and decentralized data, there has been a significant amount of recent literature developing distributed estimation or inference techniques for a variety of statistical problems [LLL13, CX14, GSS17, LLST17, SLS18, JLY19, FWWZ19, VCC19, CLZ19, LCCC20, YCC20, ZZL20, TBZ22, LSZ22, LL22, YCC22, CLLY22]. We refer readers to [GLW$^+$22] for a comprehensive review of distributed learning.

Among these works, our paper is most closely related to the literature on distributed matrix PCA. [FWWZ19] notably proposed a convenient one-shot approach that computes the top-K-dim eigenspace of the covariance matrix on each local machine and aggregates them on a central machine. They established a rigorous statistical guarantee showing that this approach achieves a sharp error rate as long as the sample size on each local machine is sufficiently large. An alternative multi-round method was simultaneously developed by [GSS17], which mainly estimates the first eigenspace by leveraging shift-and-invert preconditioning. Subsequently, [CLLY22] proposed an improved multi-round algorithm for estimating the top-K-dim eigenspace that enjoys a fast convergence rate under weaker restrictions compared to [FWWZ19]. Other works focused on providing more general statistical error bounds [ZT22] or improving the communication efficiency [HLZZ21, CBD21].

Organizations. The structure of this paper is as follows: Section 2 introduces our distributed tensor PCA algorithm for homogeneous settings and establishes its statistical error rates. Section 3 adapts the algorithm for heterogeneous environments and includes theoretical analysis. Section 4 explores the extension to transfer learning. Section 5 presents simulations and real data analyses to evaluate the performance of our methods. Some commonly used notations in this paper are defined in Section A of the supplementary material. Section B, notably, presents the asymptotic distribution of our proposed estimator, featuring distributed inference. All technical proofs and additional numerical experiments are presented in Sections C and D of the supplementary material, respectively.

2 Estimation for Homogeneous Tensors

In this section, we present the distributed tensor PCA algorithm for the homogeneous setting where each machine observes a tensor generated from the same underlying model. We first formulate the problem setup and illustrate the pooled estimator as a benchmark in Section 2.1, followed by the intuition behind its construction detailed in Section 2.2. Finally, Section 2.3 establishes theoretical guarantees on the statistical error rate for our algorithm.

2.1 Problem Setup and the Pooled Estimator

Assume a J-mode tensor $\mathcal{T}^* \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_J}$ has a Tucker decomposition

$$\mathcal{T}^* = \mathcal{G} \times_1 U_1 \times_2 U_2 \cdots \times_J U_J, \quad U_j \in \mathbb{O}^{p_j \times r_j}, \mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_J} \tag{1}$$

which decomposes $\mathcal{T}^*$ into a core tensor $\mathcal{G}$ multiplied by factor matrices $\{U_j\}$ with orthogonal columns. Suppose we observe L tensors $\{\mathcal{T}_\ell\}_{\ell=1}^L$ that are distributed on L machines and cannot be pooled together. For each machine ℓ, the tensor $\mathcal{T}_\ell$ is generated by $\mathcal{T}_\ell = \mathcal{T}^* + \mathcal{Z}_\ell$, where $\mathcal{T}^*$ is a common low-rank tensor of interest given by (1), and the noise tensor $\mathcal{Z}_\ell$ has i.i.d. normal entries

with mean zero and variance σ^2. Our goal is to estimate the singular subspace $\mathrm{Col}(U_j)$, $j \in [J]$, i.e., the linear space spanned by the columns of U_j, under the distributed setting.

The model defined in (1) is non-identifiable since it is equivalent to the model with $\widetilde{\mathcal{G}} = \mathcal{G} \times_1 O_1^\top \times_2 O_2^\top \cdots \times_J O_J^\top$ and $\widetilde{U}_j = U_j O_j$, for any $O_j \in \mathbb{O}^{r_j \times r_j}$, $j \in [J]$. However, the singular subspace $\mathrm{Col}(U_j)$ remains invariant under such orthogonal transformation and thus is identifiable. The singular subspace $\mathrm{Col}(U_j)$ can further be represented by the *projection matrix* $U_j U_j^\top$, which projects $\mathbb{R}^{p_j}$ onto $\mathrm{Col}(U_j)$ and satisfies $\widetilde{U}_j \widetilde{U}_j^\top = U_j O_j O_j^\top U_j^\top = U_j U_j^\top$. To measure the estimation error between a singular subspace spanned by $U \in \mathbb{O}^{p \times r}$ and that spanned by an estimator $\widehat{U} \in \mathbb{O}^{p \times r}$, we use a metric $\rho\big(\widehat{U}, U\big) := \big\|\widehat{U}\widehat{U}^\top - UU^\top\big\|_{\mathrm{F}}$, which is the Frobenius norm of the difference between the projection matrices of U and $\widehat{U}$. We note that ρ is equivalent to the well-known $\sin\Theta$ distance [DK70] that measures the distance between the subspaces $\mathrm{Col}(U)$ and $\mathrm{Col}\big(\widehat{U}\big)$ using principal angles, defined as

$$\big\|\sin\Theta\big(U, \widehat{U}\big)\big\|_{\mathrm{F}} = \big\|\mathrm{diag}\big(\sin(\cos^{-1}\sigma_1), \ldots, \sin(\cos^{-1}\sigma_r)\big)\big\|_{\mathrm{F}} = \sqrt{r - \sum_{i=1}^{r} \sigma_i^2}$$

where $\sigma_1, \ldots, \sigma_r$ are the singular values of $U^\top \widehat{U}$. The equivalence between ρ and $\sin\Theta$ distance can be established by

$$\rho^2\big(U, \widehat{U}\big) = \big\|UU^\top\big\|_{\mathrm{F}}^2 + \big\|\widehat{U}\widehat{U}^\top\big\|_F^2 - 2\big\|U^\top\widehat{U}\big\|_{\mathrm{F}}^2 = 2r - 2\sum_{i=1}^{r}\sigma_i^2 = 2\big\|\sin\Theta\big(U, \widehat{U}\big)\big\|_F^2$$

If the tensors $\{\mathcal{T}_\ell\}$ were allowed to be pooled onto a central machine, a standard way for estimating the singular space $\mathrm{Col}(U_j)$ would have been to conduct a Tucker decomposition on the averaged tensor $\overline{\mathcal{T}} = \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{T}_\ell = \mathcal{T}^* + \overline{\mathcal{Z}}$, where $\overline{\mathcal{Z}} = \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{Z}_\ell$ is a tensor with i.i.d. $\mathcal{N}(0, \frac{\sigma^2}{L})$ entries. Denote the estimator obtained through this method as $\widehat{U}_{\mathrm{pooled},j}$, $j \in [J]$, which serves as a benchmark against which we evaluate the effectiveness of our approach in the distributed environment.

2.2 Distributed Tensor PCA for Homogeneous Tensors

We first propose Algorithm 1 for estimating the singular subspace spanned by U_j in the distributed setting. Algorithm 1 starts with an initial estimate $\{\widehat{U}_{j,\ell}^{(0)}\}$ that can be obtained from a prototypical Tucker decomposition algorithm, for instance, higher-order SVD (HOSVD) or higher order orthogonal iteration (HOOI) [DLDMV00a, DLDMV00b], on the corresponding individual tensors $\{\mathcal{T}_\ell\}$.

Given initial estimators $\{\widehat{U}_{j,\ell}^{(0)}\}$, we obtain a local estimator for U_j by computing the left singular value matrix of $M_{j,\ell}$ on each machine ℓ, where the matrix $M_{j,\ell} \in \mathbb{R}^{p_j \times (r_1 r_2 \cdots r_J / r_j)}$ is defined as the matricization of a projected version of $\mathcal{T}_\ell$, given by

$$M_{j,\ell} = \mathcal{M}_j\big(\mathcal{T}_\ell \times_1 \widehat{U}_{1,\ell}^{(0)\top} \times_2 \widehat{U}_{2,\ell}^{(0)\top} \cdots \times_{j-1} \widehat{U}_{j-1,\ell}^{(0)\top} \times_{j+1} \widehat{U}_{j+1,\ell}^{(0)\top} \cdots \times_J \widehat{U}_{J,\ell}^{(0)\top}\big) \tag{2}$$

Under mild conditions, we show that $M_{j,\ell}$ approximately equals $U_j \mathcal{M}_j(\mathcal{G})$, and therefore, the left singular vector matrices of $\{M_{j,\ell}\}$ provide estimators $\{\widehat{U}_{j,\ell}\}$ for U_j. The local estimators $\{\widehat{U}_{j,\ell}\}$ are then sent to a central machine and further aggregated by averaging the projection matrices

Algorithm 1 Distributed Tensor PCA for Homogeneous Data

Input: Tensors distributed on local machines $\{\mathcal{T}_\ell\}$ and initial estimators $\big\{\widehat{U}_{1,\ell}^{(0)}, \widehat{U}_{2,\ell}^{(0)}, \ldots, \widehat{U}_{J,\ell}^{(0)}\big\}$, for all $\ell \in [L]$.

Output: Estimators $\big\{\widehat{U}_1, \widehat{U}_2, \ldots, \widehat{U}_J\big\}$.

for $\ell = 1, 2, \ldots, L$ **do**
  for $j = 1, 2, \ldots, J$ **do**
    Compute a local estimator $\widehat{U}_{j,\ell} = \mathrm{svd}_{r_j}(M_{j,\ell})$, where $M_{j,\ell}$ is defined in (2);
    Send $\widehat{U}_{j,\ell}$ to the central machine;
  end for
end for
for $j = 1, 2, \ldots, J$ **do**
  On the central machine, compute $\widehat{U}_j = \mathrm{svd}_{r_j}\Big[\frac{1}{L}\sum_{\ell=1}^{L} \widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\Big]$;
end for

$\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top$ over all ℓ. Finally, we compute the left singular vectors of the averaged matrix as the output estimator $\widehat{U}_j$. The communication cost of Algorithm 1 is of the order $O(\sum_{j=1}^J p_j r_j)$, which is a significant reduction from $O(\prod_{j=1}^J p_j)$, the communication cost for transferring the individual tensors themselves across machines.

Remark. In the aggregation step of Algorithm 1, we average the projection matrices $\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top$ instead of the singular vectors $\widehat{U}_{j,\ell}$ due to the non-identifiability of $\widehat{U}_{j,\ell}$, as discussed in Section 2.1. For instance, if a singular value has a multiplicity greater than 1, the corresponding singular vectors $\widehat{U}_{j,\ell}$ can be any orthonormal basis spanning the same singular subspace associated with that repeated singular value. Even if all singular values are distinctive, there is still a sign ambiguity issue, i.e., both $\widehat{U}_{j,\ell}$ and $-\widehat{U}_{j,\ell}$ may be obtained from the SVD of the same matrix, which may lead to cancellations in the averaging of $\widehat{U}_{j,\ell}$. In contrast, averaging the projection matrices avoids these issues and provides a valid estimate for the singular space spanned by U_j. Moreover, since the estimated projection matrices $\big\{\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\big\}_{\ell\in[L]}$ are not guaranteed to represent the same subspace, their average may have a rank larger than r_j. Therefore, we add an additional SVD in the final step to obtain a low-rank approximation of the averaged projection matrices, denoted as $\widehat{U}_j$, and output it as the final estimator.

Remark. In practice, the ranks $\{r_j\}_{j\in[J]}$ may be unknown and need to be specified for Algorithm 1 in the distributed way. Since the local matrix $M_{j,\ell}$ in (2) provides an approximation for $U_j\mathcal{M}_j(G)$ which has rank r_j, one may first compute a consistent rank estimator $\widehat{r}_{j,\ell}$ for $M_{j,\ell}$ using existing rank determination methods (e.g.,[CTT17, CP22, HCZ22]) and further aggregate the locally estimated ranks $\{\widehat{r}_{j,\ell}\}_{\ell\in[L]}$, for example, by averaging, to obtain a more accurate estimate for r_j. Another approach is to overparametrize each local model by specifying a conservative rank $\widehat{r}_{j,\ell} \geq r_j$, as studied in [XSCM23], and then aggregate conservatively, for example, by choosing $\widehat{r}_j$ as the maximum or certain quantile of $\{\widehat{r}_{j,\ell}\}_{\ell\in[L]}$. It is a potentially interesting future direction to investigate the performance of Algorithm 1 under overparametrization.

2.3 Theoretical Guarantee

In this section, we provide theoretical guarantees for the statistical performance of Algorithm 1. For $j \in [J]$, let Λ_j be the $r_j \times r_j$ singular value matrix of $\mathcal{M}_j(\mathcal{G})$. Let $r = \max_j r_j$ and $p = \max_j p_j$. Moreover, define $\lambda_{\max}$, $\lambda_{\min}$ to be the maximum and minimum singular value of Λ_j across all $j \in [J]$, and let $\kappa_0 = \lambda_{\max}\lambda_{\min}^{-1}$.

Theorem 2.1. *Assume that there exist constants C_1, c_1, C_2 such that, with probability at least $1 - C_1 e^{-c_1 p}$, $\big\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} - U_j U_j^\top\big\|_2 \le C_2\sqrt{p}\sigma\lambda_{\min}^{-1}$ for any $j \in [J]$ and $\ell \in [L]$. If $p_j \asymp p$ for all j, $L \lesssim p^{c_3}$ for some $c_3 > 0$, $p \gtrsim r^{J-1}$, $\kappa_0 = O(1)$, and $\lambda_{\min}/\sigma \gtrsim \sqrt{pr}$, then we have*

$$\sup_j \rho\big(\widehat{U}_j, U_j\big) = O_{\mathbb{P}}\bigg(\frac{\sigma}{\lambda_{\min}}\sqrt{\frac{pr}{L}} + \frac{pr\sigma^2}{\lambda_{\min}^2}\bigg) \tag{3}$$

where $\{\widehat{U}_j\}$ is the output of Algorithm 1

Theorem 2.1 establishes the error rate of the estimators $\big\{\widehat{U}_j\big\}$ obtained by Algorithm 1, which can be explained by a bias-variance decomposition. We take $J = 3$, $j = 1$ for an example. On each machine ℓ, the estimation error of the local estimator $\widehat{U}_{1,\ell}$ can be decomposed as

$$\widehat{U}_{1,\ell}\widehat{U}_{1,\ell}^\top - U_1U_1^\top = U_1\Lambda_1^{-2}G_1\big(U_2^\top \otimes U_3^\top\big)Z_{1,\ell}^\top U_{1\perp}U_{1\perp}^\top + U_{1\perp}U_{1\perp}^\top Z_{1,\ell}\left(U_2 \otimes U_3\right)G_1^\top\Lambda_1^{-2}U_1^\top + R_{1,\ell} \tag{4}$$

where $G_1 = \mathcal{M}_1(\mathcal{G})$, $Z_{1,\ell} = \mathcal{M}_1(\mathcal{Z}_\ell)$, $U_{1\perp}$ is the orthogonal complement of U_1, and $R_{1,\ell}$ is a remainder term. When the signal-to-noise ratio (SNR) satisfies $\lambda_{\min}/\sigma \gtrsim \sqrt{pr}$, the Frobenius norm of the first two mean-zero terms on the RHS of (4) is of the order $O\left(\sqrt{pr}\sigma\lambda_{\min}^{-1}\right)$, and the remainder term $R_{1,\ell}$ has a higher order $O\big(pr\sigma^2\lambda_{\min}^{-2}\big)$. By averaging the projection matrices on all machines, the order of the first two terms can be reduced to $O\big(\sqrt{pr}\sigma\lambda_{\min}^{-1}L^{-1/2}\big)$, while that of $R_{1,\ell}$ does not change, leading to the error rate in (3). When the SNR is sufficiently large with respect to L, the first term in (3) dominates the second one, which leads to the following corollary.

Corollary 2.1. *Under the same assumptions in Theorem 1, if we further assume that $\lambda_{\min}/\sigma \gtrsim \sqrt{prL}$, then we have*

$$\sup_j \rho\big(\widehat{U}_j, U_j\big) = O_{\mathbb{P}}\bigg(\frac{\sigma}{\lambda_{\min}}\sqrt{\frac{pr}{L}}\bigg) \tag{5}$$

Corollary 2.1 shows that, when the SNR satisfies $\lambda_{\min}/\sigma \gtrsim \sqrt{prL}$, the estimator $\widehat{U}_j$ achieves the error rate of $O(\sqrt{pr}\sigma\lambda_{\min}^{-1}L^{-1/2})$, which matches the minimax optimal lower bound. Concretely, for a tensor $\mathcal{T} = \mathcal{T}^* + \mathcal{Z}$ with $\mathcal{T}^*$ satisfying (1) and $\mathcal{Z}$ having i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, Theorem 3 in [ZX18] shows that

$$\inf_{\widehat{U}} \sup_{\mathcal{T}^* \in \mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\lambda)} \mathbb{E} r_j^{-1/2}\big\|\sin\Theta(\widehat{U}, U_j)\big\|_{\mathrm{F}} \gtrsim \Big(\frac{\sqrt{p_j}}{\lambda/\sigma} \wedge 1\Big) \tag{6}$$

where $\mathcal{F}_{\boldsymbol{p},\boldsymbol{r}}(\lambda)$ is the class of tensors with dimension $\boldsymbol{p} = (p_1, \ldots, p_J)$, rank $\boldsymbol{r} = (r_1, \ldots, r_J)$, and minimum singular value λ over all matricizations of $\mathcal{T}^*$. Recall that if we are allowed to pool all the tensors $\{\mathcal{T}_\ell\}$ on a single machine and average them, the averaged tensor satisfies $\overline{\mathcal{T}} = \mathcal{T}^* + \overline{\mathcal{Z}}$, where $\overline{\mathcal{Z}}$ has i.i.d. $\mathcal{N}(0, \frac{\sigma^2}{L})$ entries. By (6) and the equivalence between the ρ distance and $\sin\Theta$ distance, the minimax error rate that the pooled estimator $\widehat{U}_{\mathrm{pooled},j}$ (defined in Section 2.1) can achieve is the same as the rate in (5), which is the optimal rate one can expect in a non-distributed setting. Therefore, our proposed method enjoys a sharp rate when the SNR is sufficiently large.

The assumption for the initial estimators $\{\widehat{U}_{j,\ell}^{(0)}\}$ in Theorem 2.1 is consistent with Assumption 1 in [XZZ22], which can be achieved by the HOOI algorithm under a requirement that $\lambda_{\min}/\sigma \gtrsim p^{J/4}$ [ZX18]. In some scenarios, a certain number of tensors can be pooled together and averaged into a new tensor $\mathcal{T}'_\ell$ with a noise level $\sigma'_\ell < \sigma$. The requirement can be relaxed into $\lambda_{\min}/\min_\ell \sigma'_\ell \gtrsim p^{J/4}$ in such scenarios by computing the initial estimators using the locally-aggregated tensor with the smallest noise level. On the other hand, with initial estimators that satisfy the assumption in Theorem 2.1, our method only requires a weaker condition $\lambda_{\min}/\sigma \gtrsim \sqrt{prL}$ to achieve the optimal rate, as shown in Corollary 2.1.

Moreover, we do not require sample splitting for initialization, that is, the set of tensors $\{\mathcal{T}_\ell\}$ used for initialization can be the same as that used for the distributed estimation procedure in Algorithm 1. Indeed, our theoretical error bound (3) uniformly holds for all initial estimators that satisfy the assumption in Theorem 2.1, since the first two terms in decomposition (4) do not rely on the initial estimators, and the remainder term $R_{1,\ell}$ can be uniformly bounded.

3 Estimation for Heterogeneous Tensors

In this section, we generalize Algorithm 1 to a heterogeneous setting where we allow different truth tensors $\mathcal{T}^*$ on different machines. Suppose we observe L tensors $\{\mathcal{T}_\ell\}_{\ell=1}^L$ distributed on L machines, and $\mathcal{T}_\ell = \mathcal{T}_\ell^* + \mathcal{Z}_\ell$, where $\mathcal{Z}_\ell$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. Assume the truth $\mathcal{T}_\ell^* \in \mathbb{R}^{p_1\times p_2\cdots\times p_J}$ on machine ℓ satisfies a Tucker decomposition

$$\mathcal{T}_\ell^* = \mathcal{G}_\ell \times_1 [U_1\ V_{1,\ell}] \times_2 [U_2\ V_{2,\ell}] \cdots \times_J [U_J\ V_{J,\ell}], \quad U_j \in \mathbb{O}^{p_j\times r_{j,U}}, V_{j,\ell} \in \mathbb{O}^{p_j\times r_{j,V,\ell}}, \mathcal{G}_\ell \in \mathbb{R}^{r_{1,\ell}\times r_{2,\ell}\cdots\times r_{J,\ell}} \tag{7}$$

where $r_{j,\ell} = r_{j,U} + r_{j,V,\ell}$ and $U_j^\top V_{j,\ell} = 0$. For each j, the component U_j spans a common singular subspace shared by all tensors $\{\mathcal{T}_\ell\}$, while $V_{j,\ell}$ spans an individual subspace specific to each tensor. Moreover, the core $\mathcal{G}_\ell$ is allowed to be different for different ℓ. The goal is to estimate the shared singular subspace and the individual subspace separately for $j \in [J]$ and $\ell \in [L]$. Note that in this section, we assume the noise level σ^2 is identical on all machines for clear presentation. The heterogeneity on σ^2 will be further investigated in Section 4.

In addition to (7), we need a regularity condition to ensure the identifiability of the model. As discussed in Section 2.1, the model defined in (1) is non-identifiable since it is equivalent to the model with $\widetilde{\mathcal{G}} = \mathcal{G} \times_1 O_1^\top \times_2 O_2^\top \cdots \times_J O_J^\top$ and $\widetilde{U}_j = U_j O_j$, for any $O_j \in \mathbb{O}^{r_j\times r_j}$, $j \in [J]$. In the homogeneous case, the non-identifiability has no impact on estimation since the singular subspace $\mathrm{Col}(U_j)$ remains invariant under orthogonal transformation. However, under the heterogeneous setting (7), the partition of the common component U_j and the individual component $V_{j,\ell}$ is not orthogonally invariant, which necessitates identifying a fixed O_j for each j. Therefore, we require that the core tensors $\mathcal{G}_\ell$ satisfy

$$\mathcal{M}_j(\mathcal{G}_\ell)\mathcal{M}_j(\mathcal{G}_\ell)^\top = \Lambda_{j,\ell}^2 \tag{8}$$

where $\Lambda_{j,\ell}$ is a diagonal matrix with decreasing singular values for all $j \in [J]$, $\ell \in [L]$. The condition (8) is consistent with the Identification Condition (Assumption 1) in [CXCF24], which ensures the identifiability of $\big(\mathcal{G}_\ell, [U_1\ V_{1,\ell}], \ldots, [U_J\ V_{J,\ell}]\big)$ in model (7) when the singular values are distinct.

Remark. To illustrate why (8) ensures the identifiability, consider the mode-j matricization of $\widetilde{\mathcal{G}}$ under model (1), i.e., $\mathcal{M}_j\big(\widetilde{\mathcal{G}}\big) = O_j^\top G_j\big(\bigotimes_{k\neq j} O_k\big)$, where $G_j = \mathcal{M}_j(\mathcal{G})$. There exists a unique $O_j \in \mathbb{O}^{r_j\times r_j}$, the left singular value matrix of G_j, such that $\mathcal{M}_j\big(\widetilde{\mathcal{G}}\big)\mathcal{M}_j\big(\widetilde{\mathcal{G}}\big)^\top = O_j^\top G_j G_j^\top O_j$ is a diagonal matrix with distinct decreasing singular values, and thus $\widetilde{U}_j = U_j O_j$ is uniquely determined.

Algorithm 2 Distributed Tensor PCA for Heterogeneous Data

Input: Tensors distributed on local machines $\{\mathcal{T}_\ell\}$ and initial estimators $\big\{[\widehat{U}_{1,\ell}^{(0)}\ \widehat{V}_{1,\ell}^{(0)}], [\widehat{U}_{2,\ell}^{(0)}\ \widehat{V}_{2,\ell}^{(0)}], \ldots, [\widehat{U}_{J,\ell}^{(0)}\ \widehat{V}_{J,\ell}^{(0)}]\big\}$, where $\ell = 1, 2, \ldots, L$.

Output: Estimators $\big\{\widehat{U}_1, \widehat{V}_{1,\ell}, \widehat{U}_2, \widehat{V}_{2,\ell}, \ldots, \widehat{U}_J, \widehat{V}_{J,\ell}\big\}$.

for $\ell = 1, 2, \ldots, L$ **do**
  for $j = 1, 2, \ldots, J$ **do**
    Compute a local estimator $\widehat{U}_{j,\ell} = \mathrm{svd}_{r_{j,U}}\big(\widetilde{M}_{j,\ell}\big)$, where $\widetilde{M}_{j,\ell}$ is defined in (9);
    Send $\widehat{U}_{j,\ell}$ to the central machine;
  end for
end for
for $j = 1, 2, \ldots, J$ **do**
  On the central machine, compute $\widehat{U}_j = \mathrm{svd}_{r_{j,U}}\Big[\frac{1}{L}\sum_{\ell=1}^{L}\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\Big]$;
end for
Send $\widehat{U}_j$ to all machines;
for $\ell = 1, 2, \ldots, L$ **do**
  Compute $\widehat{V}_{j,\ell} = \mathrm{svd}_{r_{j,V,\ell}}\Big[\big(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\big)\widetilde{M}_{j,\ell}\Big]$;
end for

3.1 Distributed Tensor PCA for Heterogeneous Tensors

We propose Algorithm 2 for estimating the subspace spanned by U_j and $V_{j,\ell}$ under the heterogeneous setting. The initial estimators $\big\{[\widehat{U}_{j,\ell}^{(0)}\ \widehat{V}_{j,\ell}^{(0)}]\big\}$ can be obtained in the same way as the homogeneous case, for example, by the HOOI algorithm, and do not need to be partitioned into $\widehat{U}_{j,\ell}^{(0)}$ and $\widehat{V}_{j,\ell}^{(0)}$. Similar to Algorithm 1, we first obtain a local estimator $\widehat{U}_{j,\ell}$ for U_j on each machine ℓ by taking the top singular vectors of $\widetilde{M}_{j,\ell}$, where

$$\begin{aligned}\widetilde{M}_{j,\ell} &:= \mathcal{M}_j\big(\mathcal{T}_\ell \times_1 \big[\widehat{U}_{1,\ell}^{(0)}\ \widehat{V}_{1,\ell}^{(0)}\big]^\top \cdots \times_{j-1} \big[\widehat{U}_{j-1,\ell}^{(0)}\ \widehat{V}_{j-1,\ell}^{(0)}\big]^\top \times_{j+1} \big[\widehat{U}_{j+1,\ell}^{(0)}\ \widehat{V}_{j+1,\ell}^{(0)}\big]^\top \cdots \times_J \big[\widehat{U}_{J,\ell}^{(0)}\ \widehat{V}_{J,\ell}^{(0)}\big]^\top\big) \\ &\approx [U_j\ V_{j,\ell}]\mathcal{M}_j(\mathcal{G}_\ell). \end{aligned} \tag{9}$$

Then we send $\widehat{U}_{j,\ell}$ to the central machine and aggregate the projection matrices $\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top$ to compute a global estimator $\widehat{U}_j$. To estimate $\mathrm{Col}(V_{j,\ell})$, we further send $\widehat{U}_j$ back to each machine and compute the top singular vectors of $\big(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\big)\widetilde{M}_{j,\ell}$, the projection of $\widetilde{M}_{j,\ell}$ onto the orthogonal space of $\mathrm{Col}(\widehat{U}_j)$, whose top singular subspace provides a local estimator for $\mathrm{Col}(V_{j,\ell})$.

3.2 Theoretical Guarantee

In this section, we establish the statistical error rate for the estimators in Algorithm 2. For $j \in [J]$, $\ell \in [L]$, let $\Lambda_{j,\ell}$ be the $r_{j,\ell} \times r_{j,\ell}$ singular value matrix of $\mathcal{M}_j(\mathcal{G}_\ell)$ defined in (8). Define $\lambda_{\max}$, $\lambda_{\min}$ to be the maximum and minimum singular value over all $\Lambda_{j,\ell}$, respectively, and let $\kappa_0 = \lambda_{\max}\lambda_{\min}^{-1}$. Moreover, define $\Delta = \min_{j,\ell}\{\lambda_{r_{j,U},j,\ell} - \lambda_{r_{j,U}+1,j,\ell}\}$ and $\kappa = \lambda_{\max}/\Delta$, where $\lambda_{r,j,\ell}$ denotes the r-th largest singular value of $\Lambda_{j,\ell}$. Additionally, let $r = \max_{j,\ell} r_{j,\ell}$ and $r_V = \max_{j,\ell} r_{j,V,\ell}$.

Theorem 3.1. *Assume that there exist constants C_1, c_1, C_2 such that, with probability at least $1 - C_1e^{-c_1 p}$, $\big\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} + \widehat{V}_{j,\ell}^{(0)}\widehat{V}_{j,\ell}^{(0)\top} - U_jU_j^\top - V_{j,\ell}V_{j,\ell}^\top\big\|_2 \le C_2\sqrt{p}\sigma\lambda_{\min}^{-1}$ for all j, ℓ. If $p_j \asymp p$ for*

all j, $L \lesssim p^{c_3}$ for some $c_3 > 0$, $p \gtrsim r^{J-1}$, $\kappa_0 = O(1)$, $\kappa = O(1)$, and $\min(\Delta, \lambda_{\min})/\sigma \gtrsim \sqrt{pr}$, then we have

$$\sup_j \rho(\widehat{U}_j, U_j) = O_{\mathbb{P}}\left(\sqrt{\frac{pr}{L}}\frac{\sigma}{\Delta} + \frac{pr\sigma^2}{\Delta^2}\right) \tag{10}$$

and

$$\sup_{j,\ell} \rho(\widehat{V}_{j,\ell}, V_{j,\ell}) = O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}\sigma(1+\sqrt{r/L})}{\lambda_{\min}} + \frac{\sqrt{r_V}pr\sigma^2}{\lambda_{\min}^2}\right) \tag{11}$$

where $\widehat{U}_j$ and $\widehat{V}_{j,\ell}$'s are the output of Algorithm 2

The statistical rate in (10) is similar to the rate established in Theorem 2.1 in the homogeneous case with the SNR changing from $\lambda_{\min}/\sigma$ to Δ/σ. Indeed, the quantity Δ denotes the minimum gap between the minimum singular value corresponding to U_j and the maximum singular value corresponding to $V_{j,\ell}$ over all j, ℓ, which indicates the strength of the signal for estimating U_j in this heterogeneous setting and is equal to $\lambda_{\min}$ in the homogeneous setting where $V_{j,\ell} = 0$. Similar to Corollary 2.1, when SNR is sufficiently large, i.e., $\Delta/\sigma \gtrsim \sqrt{prL}$, our estimator $\widehat{U}_j$ enjoys the sharp rate $O\big(\sqrt{pr}\sigma\Delta^{-1}L^{-1/2}\big)$. For the local estimator $\widehat{V}_{j,\ell}$, the rate established in (11) matches the local rate for $[U\ V]$, $\sqrt{pr}\sigma\lambda_{\min}^{-1}$, under a mild condition that $r_V \lesssim L$ and $(\lambda_{\min}/\sigma) \gtrsim \sqrt{prr_V}$.

4 Knowledge Transfer in Distributed Tensor PCA

In this section, we explore the task of transferring knowledge from source locations to a target location within a heterogeneous setting. Knowledge Transfer seeks to enhance learning performance at a target site by leveraging insights from related source tasks. To illustrate our approach clearly, we concentrate on transferring knowledge between a single source dataset and a target dataset. However, the transfer of knowledge across multiple tasks can be managed by integrating Algorithm 2 with Algorithm 3.

4.1 Transferred Tensor PCA

Formally, suppose the source tensor $\mathcal{T}_s = \mathcal{T}_s^* + \mathcal{Z}_s$ and the target tensor $\mathcal{T}_t = \mathcal{T}_t^* + \mathcal{Z}_t$, where

$$\mathcal{T}_s^* = \mathcal{G}_s \times_1 [U_1\ V_{1,s}] \times_2 [U_2\ V_{2,s}] \cdots \times_J [U_J\ V_{J,s}], \quad U_j \in \mathbb{O}^{p_j \times r_{j,U}}, V_{j,s} \in \mathbb{O}^{p_j \times r_{j,V,s}}, \mathcal{G}_s \in \mathbb{R}^{r_{1,s} \times r_{2,s} \cdots \times r_{J,s}},$$
$$\mathcal{T}_t^* = \mathcal{G}_t \times_1 [U_1\ V_{1,t}] \times_2 [U_2\ V_{2,t}] \cdots \times_J [U_J\ V_{J,t}], \quad U_j \in \mathbb{O}^{p_j \times r_{j,U}}, V_{j,t} \in \mathbb{O}^{p_j \times r_{j,V,t}}, \mathcal{G}_t \in \mathbb{R}^{r_{1,t} \times r_{2,t} \cdots \times r_{J,t}}$$

with $U_j^\top V_{j,\ell} = 0$ and $r_{j,\ell} = r_{j,U} + r_{j,V,\ell}$ for $j \in [J]$ and $\ell = s, t$. In other words, we assume the source and target tensors share a common top-$r_{j,U}$ singular subspace spanned by U_j, but either task can have different individual components $V_{j,\ell}$. The goal is to estimate the singular subspaces spanned by U_j and $V_{j,t}$ of the target tensor $\mathcal{T}_t$. Meanwhile, we assume the noise $\mathcal{Z}_\ell$ has i.i.d. $\mathcal{N}(0, \sigma_\ell^2)$ entries for $\ell = s, t$, where the noise levels σ_s and σ_t are allowed to be different.

To achieve knowledge transfer between $\mathcal{T}_s$ and $\mathcal{T}_t$, we propose Algorithm 3, which is carefully designed for dealing with the heterogeneity in the transfer setting. Different from Algorithm 2 who treats all tensors equally, Algorithm 3 aggregates the local estimators $\widehat{U}_{j,s}$ and $\widehat{U}_{j,t}$ through a weighted average. The weights w_s and w_t are designed to optimally balance the contributions from the source and target tensors, respectively, accounting for the potential heterogeneity in their noise levels σ_s^2 and σ_t^2. The choice for the weights will be specified in the next section.

Algorithm 3 Transferred Tensor PCA

Input: Target tensor $\mathcal{T}_t$, source tensor $\mathcal{T}_s$, initial estimators $\{[\widehat{U}_{1,\ell}^{(0)}\ \widehat{V}_{1,\ell}^{(0)}], \ldots, [\widehat{U}_{J,\ell}^{(0)}\ \widehat{V}_{J,\ell}^{(0)}]\}$, and weights w_ℓ for $\ell = s, t$ that satisfy $w_s + w_t = 1$;
Output: Estimators $\{\widehat{U}_1, \widehat{V}_{1,t}, \widehat{U}_2, \widehat{V}_{2,t}, \ldots, \widehat{U}_J, \widehat{V}_{J,t}\}$.

for $\ell = s, t$ **do**
  for $j = 1, 2, \ldots, J$ **do**
    Compute a local estimator $\widehat{U}_{j,\ell} = \mathrm{svd}_{r_{j,U}}(\widetilde{M}_{j,\ell})$, where $\widetilde{M}_{j,\ell}$ is defined in (9).
    Send $\widehat{U}_{j,\ell}$ to the target machine;
  end for
end for
for $j = 1, 2, \ldots, J$ **do**
  On the target machine, compute $\widehat{U}_j = \mathrm{svd}_{r_{j,U}}\Big[w_s\widehat{U}_{j,s}\widehat{U}_{j,s}^\top + w_t\widehat{U}_{j,t}\widehat{U}_{j,t}^\top\Big]$;
  Compute $\widehat{V}_{j,t} = \mathrm{svd}_{r_{j,V,\ell}}\Big[\big(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\big)\widetilde{M}_{j,t}\Big]$;
end for

4.2 Theoretical Guarantee

Analogous to the notations in the heterogeneous settings, for $j \in [J]$ and $\ell = s, t$, let $\Lambda_{j,\ell}$ be the $r_{j,\ell} \times r_{j,\ell}$ singular value matrix of $\mathcal{M}_j(\mathcal{G}_\ell)$. Define $\lambda_{\max}$, $\lambda_{\min}$ to be the maximum and minimum singular value over all $\Lambda_{j,\ell}$, respectively, and let $\kappa_0 = \lambda_{\max}\lambda_{\min}^{-1}$. Moreover, define $\Delta = \min_{j\in[J],\ell\in\{s,t\}}\{\lambda_{r_{j,U},j,\ell} - \lambda_{r_{j,U}+1,j,\ell}\}$ and $\kappa = \lambda_{\max}/\Delta$, where $\lambda_{r,j,\ell}$ denotes the r-th largest singular value of $\Lambda_{j,\ell}$. Additionally, let $r = \max_{j,\ell} r_{j,\ell}$ and $r_V = \max_{j,\ell} r_{j,V,\ell}$.

Theorem 4.1. *Assume that there exist constants C_1, c_1, C_2 such that, with probability at least $1 - C_1e^{-c_1p}$, $\big\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} + \widehat{V}_{j,\ell}^{(0)}\widehat{V}_{j,\ell}^{(0)\top} - U_jU_j^\top - V_{j,\ell}V_{j,\ell}^\top\big\|_2 \le C_2\sqrt{p}\sigma_\ell\lambda_{\min}^{-1}$ for all $j \in [J], \ell \in \{s,t\}$. Assume $p_j \asymp p$ for all j, $p \gtrsim r^{J-1}$, $\kappa_0 = O(1)$, $\kappa = O(1)$, and $\min(\Delta, \lambda_{\min})/\max(\sigma_t, \sigma_s) \gtrsim \sqrt{pr}$. We have*

$$\sup_j \Big\|\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top\Big\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr}\sqrt{w_s^2\sigma_s^2 + w_t^2\sigma_t^2}}{\Delta}\right), \tag{12}$$

$$\sup_j \Big\|\widehat{V}_{j,t}\widehat{V}_{j,t}^\top - V_{j,t}V_{j,t}^\top\Big\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}}{\lambda_{\min}}\Big(\sigma_t + \sqrt{r}\sqrt{w_s^2\sigma_s^2 + w_t^2\sigma_t^2}\Big)\right), \tag{13}$$

where $\widehat{U}_j, \widehat{V}_{j,t}$'s are the outputs of Algorithm 3.

Theorem 4.1 establishes the statistical error rate for the estimators obtained by Algorithm 3. We note that the best rate that an estimator for the common component U_j can attain without transfer is $O(\sqrt{pr}\sigma_t\Delta^{-1})$, compared to which our transfer learning approach improves σ_t into $\sqrt{w_s^2\sigma_s^2 + w_t^2\sigma_t^2}$ if $\sigma_s \lesssim \sigma_t$. At the same time, the individual component estimator $\widehat{V}_{j,t}$ matches the local rate $O(\sqrt{pr}\sigma_t\lambda_{\min}^{-1})$ under a mild condition that $\sqrt{r_V}\sqrt{w_s^2\sigma_s^2 + w_t^2\sigma_t^2} \lesssim \sigma_t$.

Based on the error rates established in (12) and (13), we further give the optimal choice for w_s and w_t. Specifically, by minimizing $w_s^2\sigma_s^2 + w_t^2\sigma_t^2$ under the constraint $w_s + w_t = 1$, we obtain that the optimal weights are $w_s^* = \frac{\sigma_t^2}{\sigma_s^2+\sigma_t^2}$ and $w_t^* = \frac{\sigma_s^2}{\sigma_s^2+\sigma_t^2}$, leading to the error rates

$$\sup_j \Big\|\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top\Big\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr}\overline{\sigma}}{\Delta}\right) \text{ and } \sup_j \Big\|\widehat{V}_{j,t}\widehat{V}_{j,t}^\top - V_{j,t}V_{j,t}^\top\Big\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}\sigma_t}{\lambda_{\min}}\Big(1 + \frac{\sqrt{r}\overline{\sigma}}{\sigma_t}\Big)\right)$$

where $\overline{\sigma}^2 = 1/(\sigma_s^{-2} + \sigma_t^{-2})$. In practice, we can estimate σ_s and σ_t by

$$\widehat{\sigma}_\ell = \left\| \mathcal{T}_\ell - \mathcal{T}_\ell \times_1 \left[\widehat{U}_{1,\ell}\ \widehat{V}_{1,\ell}\right] \times_2 \left[\widehat{U}_{2,\ell}\ \widehat{V}_{2,\ell}\right] \cdots \times_J \left[\widehat{U}_{J,\ell}\ \widehat{V}_{J,\ell}\right] \right\|_{\mathrm{F}} / \sqrt{p_1 p_2 \cdots p_J}, \quad \ell = s, t \tag{14}$$

5 Numerical Study

In this section, we conduct numerical studies to verify the theoretical properties and evaluate the empirical performance of our proposed distributed tensor PCA algorithms. Section 5.1 presents simulation studies for both the homogeneous and heterogeneous settings. Section 5.2 then illustrates the usefulness of our algorithms on a real dataset of protein structure.

5.1 Simulations

In this section, we use Monte Carlo simulations to verify the performance of our proposed distributed Tensor PCA methods under various settings. Throughout the simulation, we consider 3-mode tensors (i.e., $J = 3$). All results are averaged over 1,000 independent runs.

Estimation for Homogeneous Tensors We first simulate the distributed homogeneous tensor setting described in Section 2.1, where

$$\mathcal{T}^* = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3, \quad U_j \in \mathbb{O}^{p_j \times r_j}, \mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$$

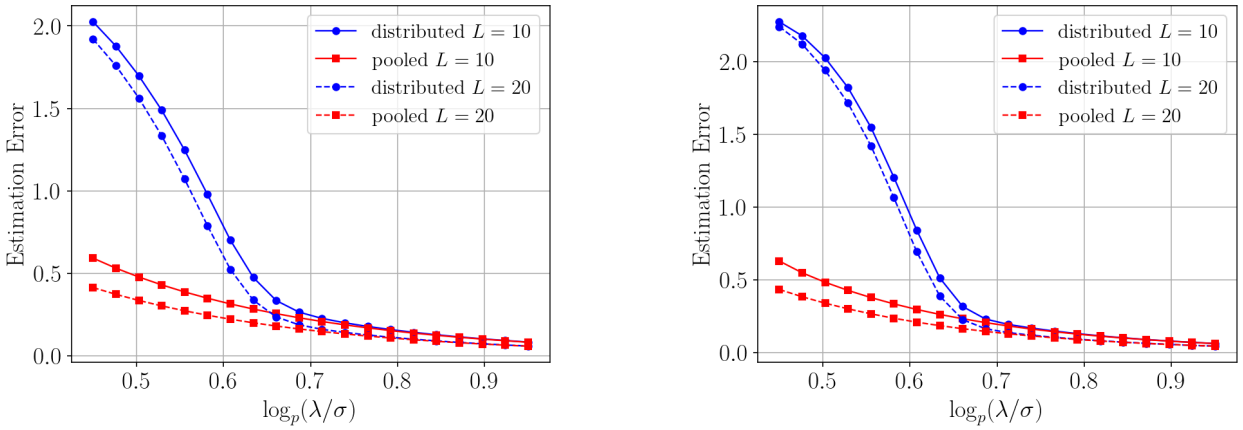
We set the dimensions $p_1 = p_2 = p_3 = p$ and the ranks $r_1 = r_2 = r_3 = r$. The core tensor $\mathcal{G}$ is generated by first sampling a tensor $\widetilde{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ with i.i.d. $\mathcal{N}(0,1)$ entries and rescaling it as $\mathcal{G} = \lambda \cdot \widetilde{\mathcal{G}} / \lambda_{\min}(\widetilde{\mathcal{G}})$, where $\lambda_{\min}(\widetilde{\mathcal{G}})$ denotes the minimum singular value of all matricizations $\mathcal{M}_j(\widetilde{\mathcal{G}})$. The generation procedure of $\mathcal{G}$ ensures that the minimum singular value of $\mathcal{G}$ is λ, which is denoted as the signal strength $\lambda_{\min}$ defined in Section 2.3. For $j = 1, 2, 3$, the matrix U_j is generated via QR decomposition on a matrix $\widetilde{U}_j \in \mathbb{R}^{p_j \times r_j}$ with i.i.d. $\mathcal{N}(0,1)$ entries. Then we independently generate $\mathcal{Z}_\ell$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and obtain $\mathcal{T}_\ell = \mathcal{T}^* + \mathcal{Z}_\ell$ for $\ell = 1, 2, \ldots, L$.

In particular, we fix $r = 3$, $\sigma = 1$, and let $p \in \{50, 100\}$, $L \in \{10, 20\}$, and $\lambda = p^\gamma$ with $\gamma \in [0.45, 0.95]$. We report the estimation error of our proposed Algorithm 1 (referred to as "distributed") for U_1, i.e., $\rho(\widehat{U}_1, U_1)$, with the SNR λ/σ ranging from $p^{0.45}$ to $p^{0.95}$. For comparison, we also report the estimation error of the pooled estimator $\widehat{U}_{\text{pooled},1}$ (referred to as "pooled") defined in Section 2.1 under the same settings. The estimation errors for $p = 50$, 100 and $L = 10$, 20 are displayed in Figure 1.

Figure 1 shows that the estimation errors of both our proposed estimator $\widehat{U}_1$ and the pooled estimator $\widehat{U}_{\text{pooled},1}$ decrease as the SNR λ/σ increases, and moreover, when the SNR is sufficiently high (e.g., $\lambda/\sigma \geq p^{0.7}$ for $p = 100$, $L = 20$), the performance of $\widehat{U}_1$ becomes indistinguishable from that of the pooled estimator, verifying that our distributed algorithm achieves the optimal minimax rate as stated in Corollary 2.1. Furthermore, we observe that increasing L from 10 to 20 leads to a noticeable reduction in the estimation error of both $\widehat{U}_1$ and the pooled estimator, consistent with the $L^{-1/2}$ rate in Theorem 2.1 and Corollary 2.1.

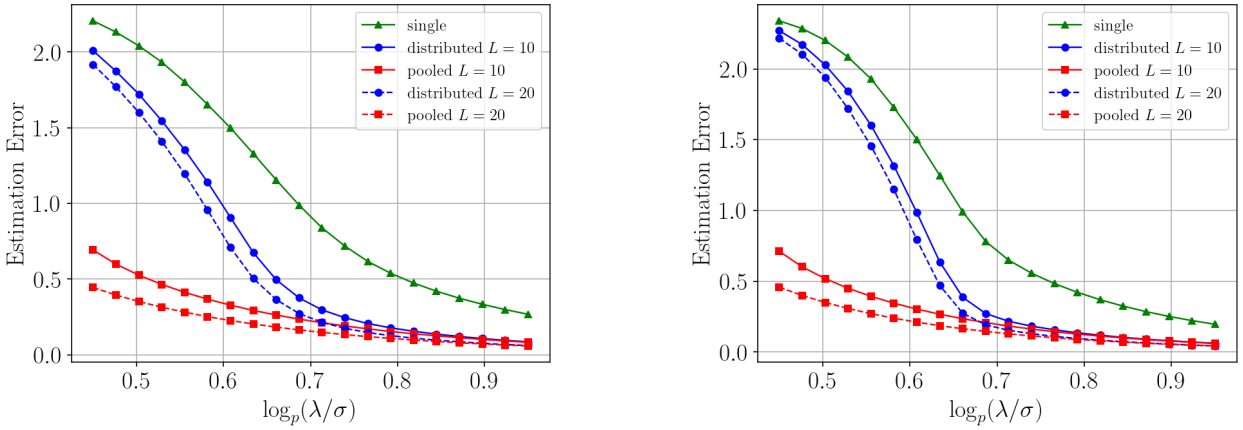
Estimation for Heterogeneous Tensors We then conduct simulations for the distributed heterogeneous setting described in Section 3, where

$$\mathcal{T}_\ell^* = \mathcal{G}_\ell \times_1 [U_1\ V_{1,\ell}] \times_2 [U_2\ V_{2,\ell}] \times_3 [U_3\ V_{3,\ell}], \quad U_j \in \mathbb{R}^{p_j \times r_{j,U}}, V_{j,\ell} \in \mathbb{R}^{p_j \times r_{j,V,\ell}}, \mathcal{G}_\ell \in \mathbb{R}^{r_{1,\ell} \times r_{2,\ell} \times r_{3,\ell}}$$



(a) $p = 50$ (b) $p = 100$

Figure 1: The estimation errors of different methods under the homogeneous setting.



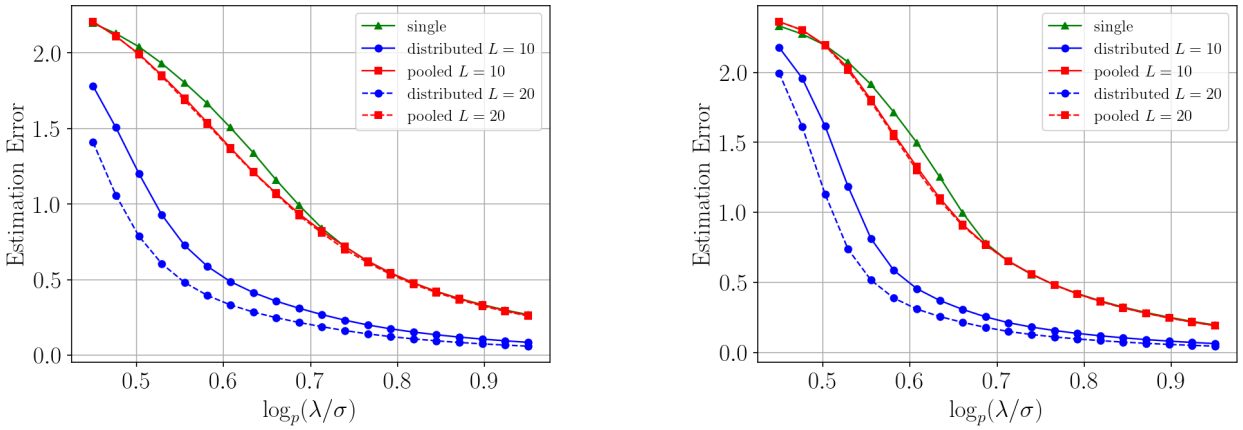
(a) $p = 50$ (b) $p = 100$

Figure 2. The estimation errors of different methods under the heterogeneous setting where the tensors share the same core.

We set the dimensions $p_j = p \in \{50, 100\}$ and the ranks $r_{j,U} = 3$, $r_{j,V,\ell} = 3$, and $r_{j,\ell} = r_{j,U} + r_{j,V,\ell} = 6$, for all $j = [3]$, $\ell \in [L]$. Similar to the homogeneous setting, the shared component U_j is generated via QR decomposition on a matrix $\widetilde{U}_j \in \mathbb{R}^{p_j \times r_{j,U}}$ with i.i.d. $\mathcal{N}(0,1)$ entries. The individual component $V_{j,\ell}$ is generated via QR decomposition on $(I_p - U_j U_j^\top)\widetilde{V}_{j,\ell}$, where $\widetilde{V}_{j,\ell} \in \mathbb{R}^{p_j \times r_{j,V,\ell}}$ has i.i.d. $\mathcal{N}(0,1)$ entries. The projection matrix $I_p - U_j U_j^\top$ ensures that $U_j^\top V_{j,\ell} = 0$. Moreover, the core tensors $\{\mathcal{G}_\ell\}$ are generated as follows.

Given a pre-specified $\lambda \in \mathbb{R}_+$, we independently sample two tensors $\widetilde{\mathcal{G}}_U \in \mathbb{R}^{3\times3\times3}$ and $\widetilde{\mathcal{G}}_V \in \mathbb{R}^{3\times3\times3}$, both with i.i.d. $\mathcal{N}(0,1)$ entries. Then we rescale them as $\mathcal{G}_U = \lambda \cdot \widetilde{\mathcal{G}}_U / \lambda_{\min}(\widetilde{\mathcal{G}}_U)$ and $\mathcal{G}_V = (\lambda/2) \cdot \widetilde{\mathcal{G}}_V / \lambda_{\max}(\widetilde{\mathcal{G}}_V)$, where $\lambda_{\min}(\lambda_{\max})(\mathcal{X})$ denotes the minimum (maximum) singular value over all matricizations $\mathcal{M}_j(\mathcal{X})$. Finally, we generate $\mathcal{G}$ as a "block-diagonal" tensor such that

$$\mathcal{G}_{i_1,i_2,i_3} = \begin{cases} (\mathcal{G}_U)_{i_1,i_2,i_3} & \text{if } 1 \le i_1, i_2, i_3 \le 3, \\ (\mathcal{G}_V)_{i_1,i_2,i_3} & \text{if } 4 \le i_1, i_2, i_3 \le 6, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$



(a) $p = 50$ (b) $p = 100$

Figure 3. The estimation errors of different methods under the heterogeneous setting where the tensors have different cores.

In other words, only the top-left $3 \times 3 \times 3$ block and the bottom-right $3 \times 3 \times 3$ block of $\mathcal{G}$ are non-zero. Furthermore, we consider two cases to generate the local core tensors $\{\mathcal{G}_\ell\}$:

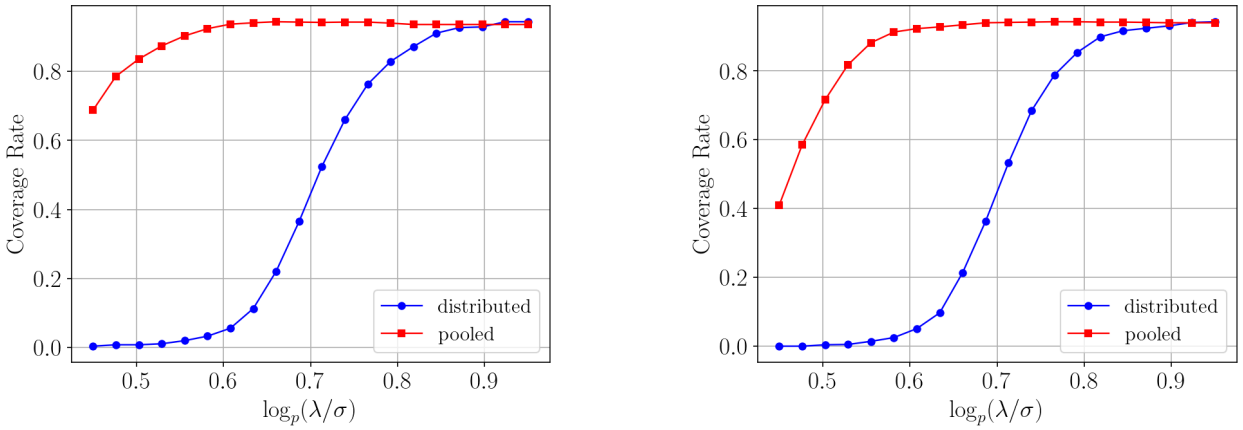
- same core: After generating $\mathcal{G}$, let $\mathcal{G}_\ell = \mathcal{G}$ for all $\ell \in [L]$;
- different cores: For each ℓ, independently generate $\mathcal{G}_\ell$ using the same procedure of generating $\mathcal{G}$ as described above.

The core tensors are constructed such that the minimum singular value gap Δ between the common and individual components equals $\lambda/2$, representing the signal strength for estimating U_j. We report the estimation errors of our proposed estimator $\widehat{U}_1$ in Algorithm 2, along with the errors of the local estimator $\widehat{U}_{1,1}$ in Algorithm 2 (referred to as "single") and the pooled estimator $\widehat{U}_{\text{pooled},1}$, for $p = 50$, 100 and $L = 10$, 20.

The results are displayed in Figures 2 and 3. In Figure 2, where all tensors share the same core tensor, our distributed estimator $\widehat{U}_1$ achieves a similar error rate as the pooled estimator when the SNR is sufficiently high, which verifies the theoretical results established in Theorem 3.1. Meanwhile, the local estimator ("single") exhibits a much higher error, highlighting the advantage of combining information across multiple tensors. In Figure 3, where the core tensors are different across machines, the performance of the pooled estimator deteriorates significantly, as the simple averaging of the tensors is invalidated due to the different cores. In contrast, our distributed estimator still achieves a decent error rate, outperforming both the pooled and local estimators. This further demonstrates the effectiveness of our method in learning the shared component in the presence of heterogeneity.

Asymptotic Distribution Furthermore, we verify the validity of the asymptotic distribution established in Section B in the supplementary material by computing the coverage rates of our proposed distributed Algorithm 4, that is, the rate that the estimated confidence region (17) covers the truth U_j. As clarified in Section B, we estimate the noise level σ by

$$\widehat{\sigma} = \left\| \mathcal{T}_1 - \mathcal{T}_1 \times_1 \widehat{U}_1 \times_2 \widehat{U}_2 \cdots \times_J \widehat{U}_J \right\|_{\mathrm{F}} / \sqrt{p_1 p_2 \cdots p_J}$$



(a) $p = 50$ (b) $p = 100$

Figure 4: The coverage rates of different methods under the heterogeneous setting.

and estimate the singular value matrix Λ_j by

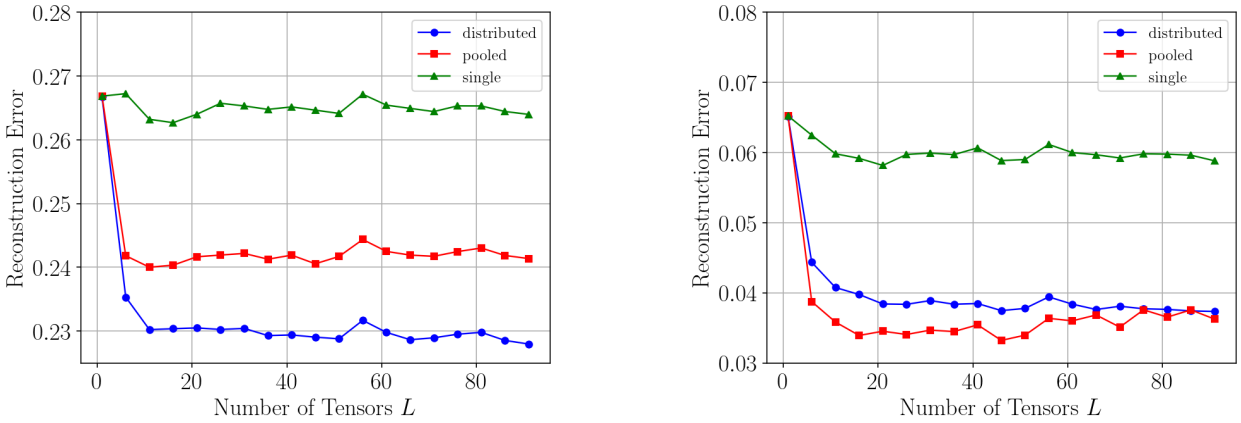
$$\widehat{\Lambda}_j = \text{the top } r_j \text{ singular values of } \mathcal{M}_j\big(\mathcal{T}_1 \times_1 \widehat{U}_1^\top \times_2 \widehat{U}_2^\top \cdots \times_{j-1} \widehat{U}_{j-1}^\top \times_{j+1} \widehat{U}_{j+1}^\top \cdots \times_J \widehat{U}_J^\top\big)$$

where $\{\widehat{U}_j\}_{j\in[J]}$ are the outputs of Algorithm 4. For comparison, we also record the coverage rates of the pooled estimator, where the confidence region for U_j is given by replacing $\widehat{U}_j$ in (17) with $\widehat{U}_{\text{pooled},j}$.

Specifically, we choose the confidence level $1-\xi$ to be 0.95 and report the coverage rates of $\widehat{U}_1$ and $\widehat{U}_{\text{pooled},1}$ with $p = 50$ and 100 in Figure 4. For both cases, our proposed estimator performs comparably to the pooled estimator when the SNR λ/σ is sufficiently large, achieving a high coverage rate around the nominal 95% level. It is worth noting that, compared to Figure 1, the requirement for SNR to achieve the asymptotic normality is more stringent than that to attain the optimal statistical error rate. This observation is consistent with our theoretical results: Corollary 2.1 guarantees the optimal statistical error rate under the condition $\lambda/\sigma \geq \sqrt{prL}$, whereas Theorem B.1, which establishes the asymptotic normality, assumes a stronger condition that $\lambda/\sigma \geq L^{1/2}(pr)^{3/4}$.

5.2 Real Data Analysis

In this section, we illustrate the usefulness of our proposed methods on the PROTEINS dataset [BOS+05, MKB+20], which contains graphs representing proteins classified as enzymes or non-enzymes. The dataset consists of 1,113 protein graphs, among which 663 are enzymes encoded as class 0, and 450 are non-enzymes encoded as class 1. Each graph represents the structure of a single protein, where the nodes represent the secondary structure elements (i.e., the helices, sheets, and turns) of the protein, and the edges connect nodes that are neighbors along the amino acid sequence or neighbors in space within the protein structure. Following the procedure in [WCC24], we employ Topological Data Analysis (TDA) to encode the topological and structural features of each graph into a three-mode tensor of dimensions $2 \times 50 \times 50$, composed of two 50×50 persistence diagrams constructed by two filtration functions. Since obtaining the ground truth U_j for real data is difficult, we evaluate the performance of different methods by the *reconstruction error* of the



(a) $\boldsymbol{r} = (1, 2, 2)$ (b) $\boldsymbol{r} = (2, 4, 4)$

Figure 5: Comparison of the reconstruction errors within class 0 of the PROTEINS dataset.

estimators. Formally, given a tensor $\mathcal{T}$, the reconstruction error of estimators $\{\widehat{U}_j\}_{j=1,2,3}$ on $\mathcal{T}$ is defined as

$$\mathrm{RE}\big(\widehat{U}_1, \widehat{U}_2, \widehat{U}_3; \mathcal{T}\big) = \frac{\Big\|\mathcal{T} - \mathcal{T} \times_1 \widehat{U}_1\widehat{U}_1^\top \times_2 \widehat{U}_2\widehat{U}_2^\top \times_3 \widehat{U}_3\widehat{U}_3^\top\Big\|_{\mathrm{F}}}{\|\mathcal{T}\|_{\mathrm{F}}}$$

which measures the difference between the original tensor and the tensor reconstructed from the estimated principal components, normalized by the Frobenius norm of the original tensor.

We first apply our proposed distributed algorithm for homogeneous data (Algorithm 1) to the tensors within class 0. Concretely, we randomly select L tensors with label 0 as the training samples, denoted by $\{\mathcal{T}_\ell\}_{\ell\in\mathcal{I}}$, where $\mathcal{I}$ is an index set with cardinality L. The training samples $\{\mathcal{T}_\ell\}_{\ell\in\mathcal{I}}$ are input into Algorithm 1 to obtain estimators $\{\widehat{U}_j\}_{j=1,2,3}$. Next, we randomly select L' tensors other than the training samples as the test set, denoted by $\{\mathcal{T}_\ell\}_{\ell\in\mathcal{I}'}$ ($\mathcal{I}\cap\mathcal{I}' = \emptyset$), and then compute the averaged reconstruction error of $\{\widehat{U}_j\}$ over the test samples, i.e., $\frac{1}{L'}\sum_{\ell\in\mathcal{I}'} \mathrm{RE}\big(\widehat{U}_1, \widehat{U}_2, \widehat{U}_3; \mathcal{T}_\ell\big)$. For comparison, we also record the reconstruction errors of two other methods:

(1) "single": the local estimators $\{\widehat{U}_{j,\ell}\}_{j=1,2,3}$ obtained using each training sample $\mathcal{T}_\ell$, $\ell\in\mathcal{I}$;

(2) "pooled": the pooled estimators obtained by decomposing the averaged tensor $\frac{1}{L}\sum_{\ell\in\mathcal{I}}\mathcal{T}_\ell$.

Specifically, we choose the ranks $\boldsymbol{r} = (r_1, r_2, r_3)$ to be $(1, 2, 2)$ or $(2, 4, 4)$, fix $L' = 100$, and let L range from 1 to 100.

Figure 5 presents the reconstruction errors of the three methods, which are averaged over 200 independent repeats of random sample selection. Note that, for the "single" method, we report the average reconstruction error over all training samples $\mathcal{T}_\ell$, that is, $\frac{1}{LL'}\sum_{\ell\in\mathcal{I}}\sum_{\ell'\in\mathcal{I}'} \mathrm{RE}\big(\widehat{U}_{1,\ell}, \widehat{U}_{2,\ell}, \widehat{U}_{3,\ell}; \mathcal{T}_{\ell'}\big)$. As shown in Figure 5, our distributed method significantly reduces the reconstruction error compared to the "single" estimators when the number of tensors in the training set is sufficiently large, which is evident for both $\boldsymbol{r} = (1, 2, 2)$ and $\boldsymbol{r} = (2, 4, 4)$. Moreover, the distributed method achieves comparable performance to the pooled method for $\boldsymbol{r} = (2, 4, 4)$ and outperforms the pooled method when $\boldsymbol{r} = (1, 2, 2)$.

6 Conclusion

This paper presents novel distributed tensor Principal Component Analysis (PCA) methods designed for both homogeneous and heterogeneous data settings. Motivated by the growing prevalence of large-scale tensors distributed across multiple locations, our work addresses the practical challenges of analyzing such data when central pooling is impractical. We develop and theoretically validate algorithms that efficiently aggregate shared low-rank subspaces and identify unique components, enhancing estimation accuracy under varying degrees of data heterogeneity.

We analyze three specific scenarios: (1) homogeneous data generated from a common noise-affected model across locations, (2) heterogeneous data generated from distinct models sharing some principal components, and (3) targeted heterogeneous settings where limited data at a location are supplemented with transferred knowledge from other sites. Our methods achieve sharp statistical guarantees, demonstrating minimax optimality while maintaining low communication costs.

Through extensive simulations and real-world data analyses, our methods show significant improvements over traditional approaches, particularly in managing data heterogeneity. Future research will explore scaling these techniques to more complex frameworks, integrating adaptive algorithms that respond to varying data characteristics and network conditions, while optimizing computational and communication efficiency in distributed environments.

References

[BOS+05] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21:i47–i56, 2005.

[BTY+21] Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu. Tensors in statistics. *Annual Review of Statistics and its Application*, 8:345–368, 2021.

[CBD21] Vasileios Charisopoulos, Austin R. Benson, and Anil Damle. Communication-efficient distributed eigenspace estimation. *SIAM Journal on Mathematics of Data Science*, 3(4):1067–1092, 2021.

[CCJ+18] Venkatesan T Chakaravarthy, Jee W Choi, Douglas J Joseph, Prakash Murali, Shivmaran S Pandian, Yogish Sabharwal, and Dheeraj Sreedhar. On optimizing distributed Tucker decomposition for ssarse tensors. In *Proceedings of the 2018 International Conference on Supercomputing*, 2018.

[CLLY22] Xi Chen, Jason D Lee, He Li, and Yun Yang. Distributed estimation for principal component analysis: An enlarged eigenspace analysis. *Journal of the American Statistical Association*, 117(540):1775–1786, 2022.

[CLZ19] Xi Chen, Weidong Liu, and Yichen Zhang. Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273, 2019.

[CP22] Anirvan Charkaborty and Victor M Panaretos. Testing for the rank of a covariance operator. *The Annals of Statistics*, 50(6):3510–3537, 2022.

[CTT17] Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, pages 2590–2617, 2017.

[CX14] Xueying Chen and Min-ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pages 1655–1684, 2014.

[CXCF24] Elynn Y Chen, Dong Xia, Chencheng Cai, and Jianqing Fan. Semi-parametric tensor factor analysis by iteratively projected singular value decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page To appear, 2024.

[DK70] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[DLDMV00a] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[DLDMV00b] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000.

[FWWZ19] Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

[GdSK+21] Bernardo B Gatto, Eulanda M dos Santos, Alessandro L Koerich, Kazuhiro Fukui, and Waldir SS Junior. Tensor analysis with n-mode generalized difference subspace. *Expert Systems with Applications*, 171:114559, 2021.

[GLW+22] Yuan Gao, Weidong Liu, Hansheng Wang, Xiaozhou Wang, Yibo Yan, and Riquan Zhang. A review of distributed statistical inference. *Statistical Theory and Related Fields*, 6(2):89–99, 2022.

[GSS17] Dan Garber, Ohad Shamir, and Nathan Srebro. Communication-efficient algorithms for distributed stochastic principal component analysis. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[HCZ22] Yuefeng Han, Rong Chen, and Cun-Hui Zhang. Rank determination in tensor factor model. *Electronic Journal of Statistics*, 16(1):1726–1803, 2022.

[HLZZ21] Zengfeng Huang, Xuemin Lin, Wenjie Zhang, and Ying Zhang. Communication-efficient distributed covariance sketch, with application to distributed PCA. *Journal of Machine Learning Research*, 22(80):1–38, 2021.

[HSS15] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

[JK20] Jun-Gi Jang and U Kang. D-Tucker: Fast and memory-efficient tucker decomposition for dense tensors. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 2020.

[JLY19] Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

[KABO10] Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, 2010.

[KB09] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[LCCC20] Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. *Journal of Machine Learning Research*, 21(180):1–51, 2020.

[LL22] Shaogao Lv and Heng Lian. Debiased distributed learning for sparse partial linear models in high dimensions. *Journal of Machine Learning Research*, 23(2):1–32, 2022.

[LLL13] Runze Li, Dennis KJ Lin, and Bing Li. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013.

[LLST17] Jason D. Lee, Qiang Liu, Yuekai Sun, and Jonathan E. Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

[LSZ22] Jiyu Luo, Qiang Sun, and Wen-Xin Zhou. Distributed adaptive Huber regression. *Computational Statistics & Data Analysis*, 169:107419, 2022.

[LXZL18] Xiaoshan Li, Da Xu, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545, 2018.

[MKB+20] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs, 2020.

[PWB20] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeir. Statistical limits of spiked tensor models. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(1):230–264, 2020.

[RM14] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[RYZ+23] Bocheng Ren, Laurence T Yang, Qingchen Zhang, Jun Feng, and Xin Nie. Blockchain-powered tensor meta-learning-driven intelligent healthcare system with IoT assistance. *IEEE Transactions on Network Science and Engineering*, 10(5):2503–2513, 2023.

[SHL21] Will Wei Sun, Botao Hao, and Lexin Li. *Tensors in Modern Statistical Learning*, pages 1–25. 2021.

[SLS18] Chengchun Shi, Wenbin Lu, and Rui Song. A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709, 2018.

[SSK16] Kijung Shin, Lee Sael, and U Kang. Fully scalable methods for distributed tensor factorization. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):100–113, 2016.

[TBZ22] Kean Ming Tan, Heather Battey, and Wen-Xin Zhou. Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of Machine Learning Research*, 23(272):1–61, 2022.

[VCC19] Stanislav Volgushev, Shih-Kang Chao, and Guang Cheng. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[WCC24] Tao Wen, Elynn Chen, and Yuzhou Chen. Tensor-view topological graph neural network. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

[WL20] Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research*, 21(154):1–38, 2020.

[Xia21] Dong Xia. Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851, 2021.

[XSCM23] Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[XZZ22] Dong Xia, Anru R. Zhang, and Yuchen Zhou. Inference for low-rank tensors – no need to debias. *The Annals of Statistics*, 50(2):1220–1245, 2022.

[YCC20] Yang Yu, Shih-Kang Chao, and Guang Cheng. Simultaneous inference for massive data: Distributed bootstrap. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[YCC22] Yang Yu, Shih-Kang Chao, and Guang Cheng. Distributed bootstrap for simultaneous inference under high dimensionality. *Journal of Machine Learning Research*, 23(195):1–77, 2022.

[YWS15] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[ZH19] Anru Zhang and Rungang Han. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, 114(528):1708–1725, 2019.

[ZT22] Runbing Zheng and Minh Tang. Limit results for distributed estimation of invariant subspaces in multiple networks inference and PCA. *arXiv preprint arXiv:2206.04306*, 2022.

[ZX18] Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

[ZZL20] Weihua Zhao, Fode Zhang, and Heng Lian. Debiasing and distributed estimation for high-dimensional quantile regression. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2569–2577, 2020.

A Notations

For any integer N, let $[N]$ denote the set $\{1, 2, \ldots, N\}$. For any positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n = O(b_n)$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a sequence of random variables X_n and a sequence of real numbers a_n, we let $X_n = O_{\mathbb{P}}(a_n)$ denote that $\{X_n/a_n\}$ is bounded in probability and $X_n = o_{\mathbb{P}}(a_n)$ denote that $\{X_n/a_n\}$ converges to zero in probability.

Define $\mathbb{O}^{p\times r} = \left\{U \in \mathbb{R}^{p\times r} \mid U^\top U = I_r\right\}$, where I_r is the $r \times r$ identity matrix. For any matrix M, denote the top-r left singular vectors of M as $\mathrm{svd}_r(M)$, and let $\mathrm{Col}(M)$ denote the linear space spanned by the columns of M. Let $\|M\|_2$ and $\|M\|_{\mathrm{F}}$ be the spectral norm and Frobenius norm of M, respectively.

For any two matrices $M_1 \in \mathbb{R}^{p_1\times q_1}$ and $M_2 \in \mathbb{R}^{p_2\times q_2}$, define $M_1 \otimes M_2 \in \mathbb{R}^{p_1p_2\times q_1q_2}$ to be their Kronecker product. For any tensor $\mathcal{X} \in \mathbb{R}^{p_1\times p_2\cdots\times p_J}$, matrix $M \in \mathbb{R}^{q_j\times p_j}$, and $j \in [J]$, let $\mathcal{X} \times_j M$ denote the mode-j matrix product of $\mathcal{X}$. That is, $(\mathcal{X} \times_j M) \in \mathbb{R}^{p_1\times\cdots\times q_j\times\cdots\times p_J}$, and

$$(\mathcal{X} \times_j M)_{i_1,i_2,\ldots,k,\ldots,i_J} = \sum_{i_j=1}^{p_j} \mathcal{X}_{i_1,i_2,\ldots,i_j,\ldots,i_J} M_{k,i_j}$$

Moreover, let $\mathcal{M}_j(\mathcal{X}) \in \mathbb{R}^{p_j\times(p_1p_2\cdots p_J/p_j)}$ denote the mode-j matricization (unfolding) of $\mathcal{X}$, which reorders the mode-j fibers of the tensor $\mathcal{X}$ to be the columns of the matrix $\mathcal{M}_j(\mathcal{X})$. Let $\lambda_{\min}(\mathcal{X})$ denote the minimum singular value over all the matricizations $\mathcal{M}_j(\mathcal{X})$, $j \in [J]$. Additionally, define $\|\mathcal{X}\|_{\mathrm{F}} = \|\mathcal{M}_1(\mathcal{X})\|_{\mathrm{F}}$.

B Statistical Inference in the Distributed Environment

In this section, we provide theoretical analysis for the asymptotic distribution of our proposed distributed method in Algorithm 1 to feature statistical inference of the singular spaces. To establish the asymptotic distribution, we develop a two-iteration distributed procedure that obtains refined local estimators using each individual tensor and then aggregates them by averaging the projection matrices. By establishing the asymptotic distribution, we provide a concise analysis of how aggregation helps to improve the statistical efficiency in the distributed environment.

The two-iteration distributed procedure is formally displayed in Algorithm 4, and we give the asymptotic distribution of the estimation error $\rho(\widehat{U}_j, U_j)$ in the following theorem.

Theorem B.1. *Assume the assumptions in Theorem 2.1 hold. Further assume that* $\lambda_{\min}/\sigma \gtrsim L^{1/2}(pr)^{3/4}$ *and* $\max(r^3, r^{J-1})/p = o(1)$. *Then we have*

$$\frac{\rho^2\big(\widehat{U}_j, U_j\big) - 2\sigma^2 L^{-1} p_j \left\|\Lambda_j^{-1}\right\|_{\mathrm{F}}^2}{\sqrt{8p_j}\sigma^2 L^{-1}\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}}} \xrightarrow{d} \mathcal{N}(0,1) \tag{16}$$

for $j \in [J]$, *where* $\widehat{U}_j$ *is the output of Algorithm 4, and* Λ_j *denotes* $r_j \times r_j$ *singular value matrix of* $\mathcal{M}_j(\mathcal{G})$

The technical reason behind the two-iteration distributed procedure is to ensure a more precise quantification of the local estimation error. Recall that in Theorem 2.1, we assume the initial estimators $\left\{\widehat{U}_{j,\ell}^{(0)}\right\}$ have an error rate of the order $O(\sqrt{p}\sigma\lambda_{\min}^{-1})$. However, to establish the asymptotic normality, we need a finer requirement for the initial estimation error, which may not be satisfied by

$\{\widehat{U}_{j,\ell}^{(0)}\}$ but is satisfied by the local estimators $\{\widehat{U}_{j,\ell}\}$ obtained in Algorithm 1. Therefore, we add one more iteration for each individual tensor in Algorithm 1 to obtain refined local estimators. This two-iteration strategy was first investigated by a recent work [XZZ22] for single tensor decomposition. In Theorem B.1, we further provide the asymptotic distribution for our distributed estimator and show that aggregation in a distributed setting helps improve the asymptotic Mean Squared Error (MSE) from a local rate $O(p^2r^2\sigma^4\lambda_{\min}^{-4})$ to the global optimal rate $O(L^{-2}p^2r^2\sigma^4\lambda_{\min}^{-4})$, when the SNR is sufficiently large.

Concretely, Theorem B.1 shows that, when the SNR satisfies $\lambda_{\min}/\sigma \gtrsim L^{1/2}(pr)^{3/4}$, the squared distance $\rho^2(\widehat{U}_j, U_j)$ has an asymptotic bias $2\sigma^2L^{-1}p_j\|\Lambda_j^{-1}\|_{\mathrm{F}}^2$ and an asymptotic standard deviation $\sqrt{8p_j}\sigma^2L^{-1}\|\Lambda_j^{-2}\|_{\mathrm{F}}$. Since $\|\Lambda_j^{-1}\|_{\mathrm{F}} \le \sqrt{r}\|\Lambda_j^{-1}\|_2 \le \sqrt{r}\lambda_{\min}^{-1}$, the asymptotic MSE of $\rho^2(\widehat{U}_j, U_j)$ is of the order $O(p^2r^2\sigma^4L^{-2}\lambda_{\min}^{-4})$, which is consistent with the error rate established in Theorem 2.1. Moreover, the pooled estimator $\widehat{U}_{\mathrm{pooled},j}$ (defined in Section 2.1) has the same asymptotic distribution as (16), indicating that our estimator achieves the optimal asymptotic MSE we can obtain in the distributed setting.

Based on Theorem B.1, we can further construct confidence regions for U_j using the proposed estimator $\widehat{U}_j$. Specifically, given a pre-specified level $1-\xi$, we construct the confidence region for U_j as follows,

$$\Big\{U \in \mathbb{O}^{p_j\times r_j} : \Big|\rho^2(\widehat{U}_j, U) - 2\widehat{\sigma}^2L^{-1}p_j\|\widehat{\Lambda}_j^{-1}\|_{\mathrm{F}}^2\Big| \le z_{1-\frac{\xi}{2}}\sqrt{8p_j}\widehat{\sigma}^2L^{-1}\|\widehat{\Lambda}_j^{-2}\|_{\mathrm{F}}\Big\} \tag{17}$$

where $z_{1-\frac{\xi}{2}}$ denotes the $\big(1-\frac{\xi}{2}\big)$-th quantile of a standard normal distribution, and $\widehat{\sigma}$ and $\widehat{\Lambda}_j$ are consistent estimators for σ and Λ_j. In practice, the noise level σ can be estimated by

$$\widehat{\sigma} = \big\|\mathcal{T}_\ell - \mathcal{T}_\ell \times_1 \widehat{U}_1 \times_2 \widehat{U}_2 \cdots \times_J \widehat{U}_J\big\|_{\mathrm{F}}/\sqrt{p_1p_2\cdots p_J}$$

and the singular value matrix Λ_j can be estimated by

$$\widehat{\Lambda}_j = \text{the top } r_j \text{ singular values of } \mathcal{M}_j\big(\mathcal{T}_\ell \times_1 \widehat{U}_1^\top \times_2 \widehat{U}_2^\top \cdots \times_{j-1} \widehat{U}_{j-1}^\top \times_{j+1} \widehat{U}_{j+1}^\top \cdots \times_J \widehat{U}_J^\top\big)$$

on any machine ℓ.

C Technical Proof of the Theoretical Results

C.1 Proof of the Results for the Homogeneous Setting

Proof for Theorem 2.1. For $j \in [J]$, $\ell \in [L]$, define $Z_{j,\ell} = \mathcal{M}_j(\mathcal{Z}_\ell)$, $T_j = \mathcal{M}_j(\mathcal{T})$, and $G_j = \mathcal{M}_j(\mathcal{G})$. Let $U_{j\perp} \in \mathbb{R}^{p_j\times(p_j-r_j)}$ be the orthogonal complement of U_j, i.e., $[U_j\ U_{j\perp}]$ is an orthogonal matrix in $\mathbb{R}^{p_j\times p_j}$. Denote the compact singular value decomposition of G_j by $U_{G_j}\Lambda_jV_{G_j}^\top$, where $U_{G_j} \in \mathbb{O}^{r_j\times r_j}$, $\Lambda_j \in \mathbb{R}^{r_j\times r_j}$, and $V_{G_j}^\top V_{G_j} = I_{r_j}$. Note the model (1) is equivalent to the model with $\widetilde{\mathcal{G}} = \mathcal{G}\times_1 O_1^\top \times_2 O_2^\top \cdots \times_J O_J^\top$ and $\widetilde{U}_j = U_jO_j$, for any $O_j \in \mathbb{O}^{r_j\times r_j}$, $j \in [J]$. Taking $O_j = U_{G_j}$ leads to $\widetilde{G}_j = \Lambda_jV_{G_j}^\top$ and thus $\widetilde{G}_j\widetilde{G}_j^\top = \Lambda_j^2$, while the projection matrix $U_jO_jO_j^\top U_j^\top = U_jU_j^\top$ remains invariant. Therefore, without loss of generality, we assume $G_jG_j^\top = \Lambda_j^2$ in the sequel. Moreover, by the definition of $\lambda_{\min}$ and κ_0, it holds that

$$\Big\|\Lambda_j^{-1}\Big\|_2 \le \lambda_{\min}^{-1}, \quad \|G_j\|_2 = \|\Lambda_j\|_2 \le \kappa_0\lambda_{\min} \tag{18}$$

Algorithm 4 Distributed Tensor PCA for Inference

Input: Tensors distributed on local machines $\{\mathcal{T}_\ell\}$ and initial estimators $\big\{\widehat{U}_{1,\ell}^{(0)}, \widehat{U}_{2,\ell}^{(0)}, \ldots, \widehat{U}_{J,\ell}^{(0)}\big\}$, for all $\ell \in [L]$.

Output: Estimators $\big\{\widehat{U}_1, \widehat{U}_2, \ldots, \widehat{U}_J\big\}$.

1: **for** $\ell = 1, 2, \ldots, L$, $j = 1, 2, \ldots, J$, $t = 1, 2$ **do**

2: Compute a local estimator

$$\widehat{U}_{j,\ell}^{(t)} = \mathrm{svd}_{r_j}\Big[\mathcal{M}_j\big(\mathcal{T}_\ell \times_1 \widehat{U}_{1,\ell}^{(t-1)\top} \times_2 \widehat{U}_{2,\ell}^{(t-1)\top} \cdots \times_{j-1} \widehat{U}_{j-1,\ell}^{(t-1)\top} \times_{j+1} \widehat{U}_{j+1,\ell}^{(t-1)\top} \cdots \times_J \widehat{U}_{J,\ell}^{(t-1)\top}\big)\Big];$$

3: **end for**

4: Send $\big\{\widehat{U}_{j,\ell}^{(2)}\big\}_{j\in[J],\ell\in[L]}$ to the central machine;

5: **for** $j = 1, 2, \ldots, J$ **do**

6: On the central machine, compute $\widehat{U}_j = \mathrm{svd}_{r_j}\Big[\frac{1}{L}\sum_{\ell=1}^{L} \widehat{U}_{j,\ell}^{(2)} \widehat{U}_{j,\ell}^{(2)\top}\Big]$;

7: **end for**

Hereafter, for clear presentation, we define $p_{-j} = \prod_{k\in[J]} p_k/p_j$, $r_{-j} = \prod_{k\in[J]} r_k/r_j$, and

$$U_{-j} = U_1 \otimes U_2 \otimes \cdots \otimes U_{j-1} \otimes U_{j+1} \otimes \cdots \otimes U_J \in \mathbb{O}^{p_{-j}\times r_{-j}}$$

For each $j \in [J]$, $\ell \in [L]$, define a "locally-good" event:

$$\begin{aligned} &E_{j,\ell}(C) \\ &:= \Bigg\{ \Big\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} - U_j U_j^\top\Big\|_2 \le C\sqrt{p}\sigma\lambda_{\min}^{-1},\ \|Z_{j,\ell}U_{-j}\|_2 \le C\sigma\sqrt{p}, \\ &\qquad \sup_{\substack{X_{j'}\in\mathbb{R}^{p_{j'}\times r_{j'}} \\ \|X_{j'}\|_2\le 1, j'\in[J]\setminus\{j\}}} \|Z_{j,\ell}(X_1 \otimes X_2 \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J)\|_2 \le C\sigma\sqrt{pr} \Bigg\} \end{aligned} \tag{19}$$

and a "globally-good" event

$$E(C) := \bigcap_{\ell=1}^{L}\bigcap_{j=1}^{J} E_{j,\ell}(C)$$

We have the following lemma for the probability of $E(C)$.

Lemma 1. *Under the assumptions in Theorem 2.1, there exist constant C_1, c_1, C_2 such that $\mathbb{P}[E(C_2)] \ge 1 - C_1 L e^{-c_1 p}$.*

Note that we assume that $L \lesssim p^{c_3}$ for some $c_3 > 0$, which implies that $E(C_2)$ has probability approaching one. For the homogeneous setting, we follow [XZZ22] to decompose $\widehat{U}_j\widehat{U}_j^\top - U_j U_j^\top$. By definition, the local estimator $\widehat{U}_{j,\ell}$ is composed of the first r_j left singular vectors of

$$\mathcal{M}_j\big(\mathcal{T}_\ell \times_1 \widehat{U}_{1,\ell}^{(0)\top} \times_2 \widehat{U}_{2,\ell}^{(0)\top} \cdots \times_{j-1} \widehat{U}_{j-1,\ell}^{(0)\top} \times_{j+1} \widehat{U}_{j+1,\ell}^{(0)\top} \cdots \times_J \widehat{U}_{J,\ell}^{(0)\top}\big) = (T_j + Z_{j,\ell})\widehat{U}_{-j}^{(0)}$$

which are also the eigenvectors of the symmetric matrix

$$(T_j + Z_{j,\ell})\widehat{U}_{-j}^{(0)}\widehat{U}_{-j}^{(0)\top}(T_j^\top + Z_{j,\ell}^\top) = T_j U_{-j} U_{-j}^\top T_j^\top + \mathfrak{E}_{j,\ell} = U_j \Lambda_j^2 U_j^\top + \mathfrak{E}_{j,\ell} \tag{20}$$

where

$$\widehat{U}_{-j}^{(0)} = \widehat{U}_1^{(0)} \otimes \widehat{U}_2^{(0)} \otimes \cdots \otimes \widehat{U}_{j-1}^{(0)} \otimes \widehat{U}_{j+1}^{(0)} \otimes \cdots \otimes \widehat{U}_J^{(0)}$$

and $\mathfrak{E}_{j,\ell}$ is a remainder term defined by

$$\begin{aligned}
\mathfrak{E}_{j,\ell} &:= \zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top + \zeta_{j,\ell,2} + \zeta_{j,\ell,2}^\top + \zeta_{j,\ell,3} + \zeta_{j,\ell,4} + \zeta_{j,\ell,5}, \\
\zeta_{j,\ell,1} &:= T_j U_{-j} U_{-j}^\top Z_{j,\ell}^\top, \\
\zeta_{j,\ell,2} &:= T_j \left[\widehat{U}_{-j}^{(0)} \widehat{U}_{-j}^{(0)\top} - U_{-j} U_{-j}^\top\right] Z_{j,\ell}^\top, \\
\zeta_{j,\ell,3} &:= Z_{j,\ell} U_{-j} U_{-j}^\top Z_{j,\ell}^\top, \\
\zeta_{j,\ell,4} &:= Z_{j,\ell} \left[\widehat{U}_{-j}^{(0)} \widehat{U}_{-j}^{(0)\top} - U_{-j} U_{-j}^\top\right] Z_{j,\ell}^\top, \\
\zeta_{j,\ell,5} &:= T_j \left[\widehat{U}_{-j}^{(0)} \widehat{U}_{-j}^{(0)\top} - U_{-j} U_{-j}^\top\right] T_j^\top
\end{aligned} \tag{21}$$

The last equality of (20) follows from the facts that $T_j = U_j G_j U_{-j}^\top$ and $G_j G_j^\top = \Lambda_j^2$. We use the following lemma to provide upper bounds for each term in (21).

Lemma 2. *Under the event $E(C_2)$ and the assumptions in Theorem 2.1, there exists some absolute constant $C_2' > 0$ such that*

$$\begin{aligned}
&\|\zeta_{j,\ell,1}\|_2 \le C_2' \kappa_0 \lambda_{\min} \sigma \sqrt{p}, \quad \|\zeta_{j,\ell,2}\|_2 \le C_2' \kappa_0 p \sigma^2 \sqrt{r}, \\
&\|\zeta_{j,\ell,3}\|_2 \le C_2' p \sigma^2, \quad \|\zeta_{j,\ell,4}\|_2 \le C_2' p^{3/2} \sqrt{r} \sigma^3 \lambda_{\min}^{-1}, \quad \|\zeta_{j,\ell,5}\|_2 \le C_2' \kappa_0^2 p \sigma^2
\end{aligned} \tag{22}$$

By Lemma 2, under the event $E(C_2)$, $\|\mathfrak{E}_{j,\ell}\|_2 \lesssim \kappa_0 \lambda_{\min} \sigma \sqrt{p} < \lambda_{\min}^2/2$. Applying Theorem 1 in [Xia21] yields

$$\begin{aligned}
\widehat{U}_{j,\ell} \widehat{U}_{j,\ell}^\top - U_j U_j^\top =\ & U_j \Lambda_j^{-2} U_j^\top \mathfrak{E}_{j,\ell} U_{j\perp} U_{j\perp}^\top + U_{j\perp} U_{j\perp}^\top \mathfrak{E}_{j,\ell} U_j \Lambda_j^{-2} U_j^\top \\
& - U_j \Lambda_j^{-2} U_j^\top \mathfrak{E}_{j,\ell} U_{j\perp} U_{j\perp}^\top \mathfrak{E}_{j,\ell} U_j \Lambda_j^{-2} U_j^\top - U_j \Lambda_j^{-2} U_j^\top \mathfrak{E}_{j,\ell} U_j \Lambda_j^{-2} U_j^\top \mathfrak{E}_{j,\ell} U_{j\perp} U_{j\perp}^\top \\
& - U_{j\perp} U_{j\perp}^\top \mathfrak{E}_{j,\ell}^{(1)} U_j \Lambda_j^{-2} U_j^\top \mathfrak{E}_{j,\ell} U_j \Lambda_j^{-2} U_j^\top + \mathfrak{R}_{j,\ell}
\end{aligned} \tag{23}$$

where $\|\mathfrak{R}_{j,\ell}\|_2 \lesssim \kappa_0^3 \sigma^3 p^{3/2} / \lambda_{\min}^3$. Then plugging (21) into (23) and using the fact that $U_{j\perp}^\top T_j = 0$, we obtain

$$\widehat{U}_{j,\ell} \widehat{U}_{j,\ell}^\top - U_j U_j^\top = \mathfrak{S}_{j,\ell,1} + \mathfrak{S}_{j,\ell,2} + \mathfrak{S}_{j,\ell,3} \tag{24}$$

where for $j \in [J]$ and $\ell \in [L]$,

$$\begin{aligned}
\mathfrak{S}_{j,\ell,1} &:= U_j \Lambda_j^{-2} U_j^\top \zeta_{j,\ell,1} U_{j\perp} U_{j\perp}^\top + U_{j\perp} U_{j\perp}^\top \zeta_{j,\ell,1}^\top U_j \Lambda_j^{-2} U_j^\top \\
&\quad = U_j \Lambda_j^{-2} G_j U_{-j}^\top Z_{j,\ell}^\top U_{j\perp} U_{j\perp}^\top + U_{j\perp} U_{j\perp}^\top Z_{j,\ell} U_{-j} G_j^\top \Lambda_j^{-2} U_j^\top, \\
\mathfrak{S}_{j,\ell,2} &:= U_j \Lambda_j^{-2} U_j^\top \zeta_{j,\ell,2}^{(t)} U_{j\perp} U_{j\perp}^\top + U_{j\perp} U_{j\perp}^\top \zeta_{j,\ell,2}^\top U_j \Lambda_j^{-2} U_j^\top \\
&\quad + U_j \Lambda_j^{-2} U_j^\top \zeta_{j,\ell,3} U_{j\perp} U_{j\perp}^\top + U_{j\perp} U_{j\perp}^\top \zeta_{j,\ell,3} U_j \Lambda_j^{-2} U_j^\top \\
&\quad - U_j \Lambda_j^{-2} U_j^\top \zeta_{j,\ell,1} U_{j\perp} U_{j\perp}^\top \zeta_{j,\ell,1}^\top U_j \Lambda_j^{-2} U_j^\top \\
&\quad - U_j \Lambda_j^{-2} U_j^\top (\zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top) U_j \Lambda_j^{-2} U_j^\top \zeta_{j,\ell,1} U_{j\perp} U_{j\perp}^\top \\
&\quad - U_{j\perp} U_{j\perp}^\top \zeta_{j,\ell,1}^\top U_j \Lambda_j^{-2} U_j^\top (\zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top) U_j \Lambda_j^{-2} U_j^\top
\end{aligned} \tag{25}$$

and

$$\begin{aligned}\mathfrak{S}_{j,\ell,3} :=\ & U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,4}U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,4}U_j\Lambda_j^{-2}U_j^\top \\ & - U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}-\zeta_{j,\ell,1}-\zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}U_{j\perp}U_{j\perp}^\top \\ & - U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}-\zeta_{j,\ell,1}-\zeta_{j,\ell,1}^\top\right]U_{j\perp}U_{j\perp}^\top \\ & - U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}-\zeta_{j,\ell,1}-\zeta_{j,\ell,1}^\top\right]U_{j\perp}U_{j\perp}^\top\mathfrak{E}_{j,\ell}U_j\Lambda_j^{-2}U_j^\top \\ & - U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}U_{j\perp}U_{j\perp}^\top\left[\mathfrak{E}_{j,\ell}-\zeta_{j,\ell,1}-\zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top \\ & - U_{j\perp}U_{j\perp}^\top\left[\mathfrak{E}_{j,\ell}-\zeta_{j,\ell,1}-\zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}U_j\Lambda_j^{-2}U_j^\top \\ & - U_{j\perp}U_{j\perp}^\top\mathfrak{E}_{j,\ell}U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}-\zeta_{j,\ell,1}-\zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top + \mathfrak{R}_{j,\ell}\end{aligned} \tag{26}$$

Under $E(C_2)$, it holds that $\|Z_{j,\ell}U_{-j}\|_2 \le C_2\sigma\sqrt{p}$. Using the fact that $\left\|\Lambda_j^{-1}G_j\right\|_2 = 1$ and $\left\|\Lambda_j^{-1}\right\|_2 \le \lambda_{\min}^{-1}$, we have

$$\|\mathfrak{S}_{j,\ell,1}\|_2 = \left\|U_j\Lambda_j^{-2}G_jU_{-j}^\top Z_{j,\ell}^\top U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top Z_{j,\ell}U_{-j}G_j^\top\Lambda_j^{-2}U_j^\top\right\|_2 \lesssim \lambda_{\min}^{-1}\sigma\sqrt{p} \tag{27}$$

with probability at least $1 - C_1e^{-c_1p}$. By (22), it holds that

$$\|\mathfrak{S}_{j,\ell,2}\|_2 \le C_3\kappa_0^2pr^{1/2}\sigma^2\lambda_{\min}^{-2}, \quad \|\mathfrak{S}_{j,\ell,3}\|_2 \le C_3\kappa_0^3p^{3/2}r^{1/2}\sigma^3\lambda_{\min}^{-3} \tag{28}$$

for some absolute constant $C_3 > 0$. Then

$$\begin{aligned}&\frac{1}{L}\sum_{\ell=1}^L\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top - U_jU_j^\top \\ &= \overline{\mathfrak{S}_{j,1}} + \overline{\mathfrak{S}_{j,2}} + \overline{\mathfrak{S}_{j,3}} \\ &= U_j\Lambda_j^{-2}G_jU_{-j}^\top\overline{Z}_j^\top U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\Lambda_j^{-2}U_j^\top + \overline{\mathfrak{S}_{j,2}} + \overline{\mathfrak{S}_{j,3}}\end{aligned} \tag{29}$$

where $\overline{Z}_j := (1/L)\sum_{\ell=1}^L Z_{j,\ell}$ and $\overline{\mathfrak{S}_{j,k}} := (1/L)\sum_{\ell=1}^L\mathfrak{S}_{j,\ell,k}$, $k \in \{1,2,3\}$. Note that

$$\left\|\overline{\mathfrak{S}_{j,2}}\right\|_2 \lesssim \kappa_0^2pr^{1/2}\sigma^2\lambda_{\min}^{-2}, \quad \left\|\overline{\mathfrak{S}_{j,3}}\right\|_2 \lesssim \kappa_0^3p^{3/2}r^{1/2}\sigma^3\lambda_{\min}^{-3}$$

Since $Z_{j,\ell}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, the averaged noise $\overline{Z}_j$ has i.i.d. $\mathcal{N}\left(0,\frac{\sigma^2}{L}\right)$ entries. By the proof of Lemma 1, it holds that $\left\|\overline{Z}_jU_{-j}\right\|_2 \lesssim \sigma\sqrt{p/L}$ and thus

$$\left\|\overline{\mathfrak{S}_{j,1}}\right\|_2 = \left\|U_j\Lambda_j^{-2}G_jU_{-j}^\top\overline{Z}_j^\top U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\Lambda_j^{-2}U_j^\top\right\|_2 \lesssim \lambda_{\min}^{-1}\sigma\sqrt{p/L} \tag{30}$$

with probability at least $1 - C_1e^{-c_1p}$.

Therefore, we obtain that

$$\frac{1}{L}\sum_{\ell=1}^L\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top - U_jU_j^\top = \overline{\mathfrak{S}_{j,1}} + \overline{\mathfrak{S}_{j,2}} + \overline{\mathfrak{S}_{j,3}} \tag{31}$$

with

$$\left\|\overline{\mathfrak{S}_{j,1}}\right\|_2 \lesssim \lambda_{\min}^{-1}\sigma\sqrt{p/L}, \quad \left\|\overline{\mathfrak{S}_{j,2}}\right\|_2 \lesssim \kappa_0^2pr^{1/2}\sigma^2\lambda_{\min}^{-2}, \quad \left\|\overline{\mathfrak{S}_{j,3}}\right\|_2 \lesssim \kappa_0^3p^{3/2}r^{1/2}\sigma^3\lambda_{\min}^{-3} \tag{32}$$

with probability at least $1 - C_1' L e^{-c_1' p}$ for some absolute constant $C_1', c_1' > 0$.

Since the columns of $\widehat{U}_j$ are the first r_j eigenvectors of $\frac{1}{L}\sum_{\ell=1}^{L} \widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top$, similar to (23),

$$\begin{aligned}\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top = {} & U_jU_j^\top \left(\overline{\mathfrak{S}_{j,1}} + \overline{\mathfrak{S}_{j,2}}\right) U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top \left(\overline{\mathfrak{S}_{j,1}} + \overline{\mathfrak{S}_{j,2}}\right) U_jU_j^\top \\ & - U_jU_j^\top \overline{\mathfrak{S}_{j,1}} U_{j\perp}U_{j\perp}^\top \overline{\mathfrak{S}_{j,1}} U_jU_j^\top + \widetilde{\mathfrak{R}}_j\end{aligned} \tag{33}$$

where $\left\|\widetilde{\mathfrak{R}}_j\right\|_2 \lesssim \kappa_0^3 p^{3/2} r^{1/2} \sigma^3 \lambda_{\min}^{-3}$. By (32), we obtain that

$$\left\|\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top\right\|_{\mathrm{F}} \le \sqrt{r}\left\|\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top\right\|_2 = O_{\mathbb{P}}\left(\lambda_{\min}^{-1}\sigma\sqrt{pr/L} + \kappa_0^2 pr\sigma^2\lambda_{\min}^{-2}\right)$$

which proves (3). □

C.2 Proof of the Results for the Heterogeneous Setting

Proof of Theorem 3.1. Similar to the proof of Theorem 2.1, define $T_{j,\ell}^* = \mathcal{M}_j(\mathcal{T}_\ell^*)$, and

$$[UV]_{-j,\ell} = [U_1\, V_{1,\ell}] \otimes [U_2\, V_{2,\ell}] \otimes \cdots \otimes [U_{j-1}\, V_{j-1,\ell}] \otimes [U_{j+1}\, V_{j+1,\ell}] \otimes \cdots \otimes [U_J\, V_{J,\ell}]$$

Recall that $\Lambda_{j,\ell}$ denotes the singular value matrix of $\mathcal{M}_j(\mathcal{G}_\ell)$. Let $\Lambda_{j,U}$ denote the top-left $r_{j,U} \times r_{j,U}$ diagonal block of $\Lambda_{j,\ell}$ and Λ_{j,V_ℓ} be the remaining block, i.e., $\Lambda_{j,\ell} = \begin{pmatrix}\Lambda_{j,U} & 0 \\ 0 & \Lambda_{j,V_\ell}\end{pmatrix}$. By (8) and (20), the columns of $\widehat{U}_{j,\ell}$ are the first $r_{j,U}$ eigenvectors of

$$U_j\Lambda_{j,U}^2U_j^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top + \mathfrak{E}_{j,\ell} \tag{34}$$

where $\mathfrak{E}_{j,\ell}$ is a remainder term satisfying

$$\sup_{j,\ell}\left\|\mathfrak{E}_{j,\ell}\right\|_2 \le C_2\lambda_{\min}\kappa_0\sigma\sqrt{p} \tag{35}$$

$$\sup_{j,\ell}\left\|\mathfrak{E}_{j,\ell} - \zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top\right\|_2 \le C_2\kappa_0^2 p\sigma^2\sqrt{r} \tag{36}$$

for

$$\zeta_{j,\ell,1} = T_{j,\ell}^*[UV]_{-j,\ell}[UV]_{-j,\ell}^\top Z_{j,\ell}^\top$$

and some absolute constant C_2, with probability tending to one.

In the following proof, the Davis-Kahan Theorem refers to the variant provided in [YWS15]. By Davis-Kahan Theorem, it holds that

$$\sup_{j,\ell}\left\|\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top - U_jU_j^\top\right\|_{\mathrm{F}} \lesssim \sup_{j,\ell}\frac{\sqrt{r_{j,U}}\left\|\mathfrak{E}_{j,\ell}\right\|_2}{\lambda_{r_{j,U},j,\ell}^2 - \lambda_{r_{j,U}+1,j,\ell}^2} \lesssim \frac{\sqrt{pr_U}\kappa\sigma}{\Delta} \tag{37}$$

where $r_U = \max_j r_{j,U}$.

Define

$$\Sigma_j^* = \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]$$

and let $U_j^* = \mathrm{svd}_{r_{j,U}}(\Sigma_j^*)$. By (37), we have

$$\sup_{j,\ell}\left\|\mathbb{E}\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right] - U_jU_j^\top\right\|_{\mathrm{F}} \lesssim \frac{\sqrt{pr_U}\kappa\sigma}{\Delta} \tag{38}$$

and thus

$$\left\|\Sigma_j^* - U_j U_j^\top\right\|_2 \le \left\|\Sigma_j^* - U_j U_j^\top\right\|_{\mathrm{F}} \le \frac{\sqrt{pr_U}\kappa\sigma}{\Delta}$$

Under the assumption that $(\Delta/\sigma) \gtrsim \kappa\sqrt{pr}$, by Weyl's inequality, we obtain that $\lambda_{r_{j,U}}(\Sigma_j^*) - \lambda_{r_{j,U}+1}(\Sigma_j^*) = 1 + o(1)$, and hence, by Davis-Kahan Theorem,

$$\left\|\widehat{U}_j\widehat{U}_j^\top - U_j^* U_j^{*\top}\right\|_{\mathrm{F}} \lesssim \left\|\frac{1}{L}\sum_{\ell=1}^{L}\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top - \Sigma_j^*\right\|_{\mathrm{F}} \le \frac{1}{L}\left\|\sum_{\ell=1}^{L}(1-\mathbb{E})\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]\right\|_{\mathrm{F}}$$

By the proof of Lemma 4 in [FWWZ19], we have that

$$\left\|\sum_{\ell=1}^{L}(1-\mathbb{E})\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]\right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\sqrt{L\sup_{\ell}\left\|(1-\mathbb{E})\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]\right\|_{\mathrm{F}}^2}\right) = O_{\mathbb{P}}\left(\frac{\sqrt{Lpr_U}\kappa\sigma}{\Delta}\right)$$

where the second equality follows from (37) and (38). Therefore, we obtain that

$$\left\|\widehat{U}_j\widehat{U}_j^\top - U_j^* U_j^{*\top}\right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr_U}\kappa\sigma}{\sqrt{L}\Delta}\right) \tag{39}$$

Now we focus on bounding $\left\|U_j^* U_j^{*\top} - U_j U_j^\top\right\|_{\mathrm{F}}$. Let the $(\boldsymbol{u}_{1,j,\ell}, \ldots, \boldsymbol{u}_{p_j,j,\ell})$ be the eigenvectors of $U_j\Lambda_{j,U}^2 U_j^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2 V_{j,\ell}^\top$ and the corresponding eigenvalues be $\lambda_{1,j,\ell}^2 > \cdots > \lambda_{p_j,j,\ell}^2$. For $k = 1, \ldots, r_{j,U}$, let $O_{k,j,\ell} = \sum_{k'=r_{j,U}+1}^{+\infty}(\lambda_{k',j,\ell}^2 - \lambda_{k,j,\ell}^2)^{-1}\boldsymbol{u}_{k'}\boldsymbol{u}_{k'}^\top$, and define

$$f_{j,\ell} : \mathbb{R}^{p_j \times r_{j,U}} \to \mathbb{R}^{p_j \times r_{j,U}}, (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{r_{j,U}}) \to (-O_{1,j,\ell}\boldsymbol{w}_1, \ldots, -O_{r_{j,U},j,\ell}\boldsymbol{w}_{r_{j,U}})$$

Since

$$\sup_{j,\ell}\|\mathfrak{E}_{j,\ell}\|_2 \lesssim \lambda_{\max}\sigma\sqrt{p} \lesssim \Delta^2 \le \lambda_{r_{j,U},j,\ell}^2 - \lambda_{r_{j,U}+1,j,\ell}^2$$

(using $(\Delta/\sigma) \gtrsim \kappa\sqrt{pr_{j,U}}$), we can apply Lemma 2 in [FWWZ19] and obtain that

$$\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top = U_j U_j^\top + f_{j,\ell}(\mathfrak{E}_{j,\ell}U_j)U_j^\top + U_j f_{j,\ell}(\mathfrak{E}_{j,\ell}U_j)^\top + \mathfrak{R}_{j,\ell}$$

where the remainder term satisfies

$$\sup_{j,\ell}\|\mathfrak{R}_{j,\ell}\|_{\mathrm{F}} \lesssim \frac{\sqrt{r_{j,U}}\,\|\mathfrak{E}_{j,\ell}\|_2^2}{\Delta^4} \lesssim \frac{\sqrt{r_U}p\kappa^2\sigma^2}{\Delta^2}$$

By (36) and $\mathbb{E}[\zeta_{j,\ell,1}] = 0$, we have that $\|\mathbb{E}[\mathfrak{E}_{j,\ell}]\|_2 \lesssim \kappa_0^2 p\sigma^2\sqrt{r}$, and hence

$$\left\|\mathbb{E}\left[f_{j,\ell}(\mathfrak{E}_{j,\ell}U_j)U_j^\top\right]\right\|_{\mathrm{F}} = \|f_{j,\ell}(\mathbb{E}[\mathfrak{E}_{j,\ell}U_j])\|_{\mathrm{F}} \le \Delta^{-2}\|\mathbb{E}[\mathfrak{E}_{j,\ell}]U_j\|_{\mathrm{F}} \le \Delta^{-2}\sqrt{r_{j,U}}\,\|\mathbb{E}[\mathfrak{E}_{j,\ell}]\|_2 \lesssim \kappa_0^2 pr\sigma^2\Delta^{-2}$$

Therefore, we obtain

$$\left\|\mathbb{E}\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right] - U_j U_j^\top\right\|_{\mathrm{F}} \lesssim \frac{pr(\kappa^2 + \kappa_0^2)\sigma^2}{\Delta^2}$$

which implies that

$$\left\|U_j^* U_j^{*\top} - U_j U_j^\top\right\|_{\mathrm{F}} \lesssim \frac{pr(\kappa^2 + \kappa_0^2)\sigma^2}{\Delta^2}$$

Together with (39), we conclude that

$$\left\|\widehat{U}_j\widehat{U}_j^\top - U_jU_j^\top\right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\sqrt{\frac{pr}{L}}\frac{\kappa\sigma}{\Delta} + \frac{pr(\kappa^2+\kappa_0^2)\sigma^2}{\Delta^2}\right)$$

For $\widehat{V}_{j,\ell}$, note that its columns are the eigenvectors of

$$\begin{aligned}
&\left(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\right)\left(U_j\Lambda_{j,U}^2U_j^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top + \mathfrak{E}_{j,\ell}\right)\left(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\right)\\
&= \left(I_{p_j} - U_jU_j^\top + U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)\left(U_j\Lambda_{j,U}^2U_j^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top + \mathfrak{E}_{j,\ell}\right)\left(I_{p_j} - U_jU_j^\top + U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)\\
&= V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top\left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right) + \left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top\\
&\quad + \left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)\left(U_j\Lambda_{j,U}^2U_j^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top\right)\left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right) + \left(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\right)\mathfrak{E}_{j,\ell}\left(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\right)
\end{aligned}$$

Since

$$\sup_{j,\ell}\left\|V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top\left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{pr}{L}}\frac{\lambda_{\max}^2\kappa\sigma}{\Delta} + \frac{pr\lambda_{\max}^2(\kappa^2+\kappa_0^2)\sigma^2}{\Delta^2}\right)$$

$$\begin{aligned}
&\sup_{j,\ell}\left\|\left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)\left(U_j\Lambda_{j,U}^2U_j^\top + V_{j,\ell}\Lambda_{j,V_\ell}^2V_{j,\ell}^\top\right)\left(U_jU_j^\top - \widehat{U}_j\widehat{U}_j^\top\right)\right\|_2\\
&= O_{\mathbb{P}}\left(\frac{pr}{L}\frac{\lambda_{\max}^2\kappa^2\sigma^2}{\Delta^2} + \frac{p^2r^2\lambda_{\max}^2(\kappa^2+\kappa_0^2)^2\sigma^4}{\Delta^4}\right)\\
&= O_{\mathbb{P}}\left(\sqrt{\frac{pr}{L}}\frac{\lambda_{\max}^2\kappa\sigma}{\Delta} + \frac{pr\lambda_{\max}^2(\kappa^2+\kappa_0^2)\sigma^2}{\Delta^2}\right)
\end{aligned}$$

and

$$\sup_{j,\ell}\left\|\left(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\right)\mathfrak{E}_{j,\ell}\left(I_{p_j} - \widehat{U}_j\widehat{U}_j^\top\right)\right\|_2 = O_{\mathbb{P}}\left(\sqrt{p}\lambda_{\max}\sigma\right)$$

By Davis-Kahan Theorem, we obtain that

$$\begin{aligned}
\sup_{j,\ell}\left\|\widehat{V}_{j,\ell}\widehat{V}_{j,\ell}^\top - V_{j,\ell}V_{j,\ell}^\top\right\|_{\mathrm{F}} &= O_{\mathbb{P}}\left(\frac{\sqrt{pr_{j,V,\ell}}\lambda_{\max}\sigma + \sqrt{prr_{j,V,\ell}/L}\lambda_{\max}^2\kappa\sigma/\Delta + \frac{\sqrt{r_{j,V,\ell}}pr\lambda_{\max}^2(\kappa^2+\kappa_0^2)\sigma^2}{\Delta^2}}{\lambda_{\min}^2}\right)\\
&= O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}\kappa_0\sigma(1+\kappa^2\sqrt{r/L})}{\lambda_{\min}} + \frac{\sqrt{r_V}pr\kappa^2(\kappa^2+\kappa_0^2)\sigma^2}{\lambda_{\min}^2}\right)
\end{aligned}$$

□

Proof of Theorem 4.1. Define $\Sigma_j^* = \sum_{\ell=s,t} w_\ell\mathbb{E}\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]$ and let $U_j^* = \mathrm{svd}_{r_{j,U}}(\Sigma_j^*)$. By the proof of Theorem 3.1, we have that

$$\begin{aligned}
\sup_j\left\|\widehat{U}_j\widehat{U}_j^\top - U_j^*U_j^{*\top}\right\|_{\mathrm{F}} &\lesssim \left\|\sum_{\ell=s,t} w_\ell(1-\mathbb{E})\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]\right\|_{\mathrm{F}}\\
&= O_{\mathbb{P}}\left(\sqrt{\sum_{\ell=s,t} w_\ell^2\left\|(1-\mathbb{E})\left[\widehat{U}_{j,\ell}\widehat{U}_{j,\ell}^\top\right]\right\|_{\mathrm{F}}^2}\right)\\
&= O_{\mathbb{P}}\left(\frac{\sqrt{pr_U}\kappa\sqrt{w_s^2\sigma_s^2 + w_t^2\sigma_t^2}}{\Delta}\right)
\end{aligned}$$

and

$$\sup_j \left\| U_j^* U_j^{*\top} - U_j U_j^\top \right\|_{\mathrm{F}} \lesssim \frac{pr(\kappa^2+\kappa_0^2)(w_s\sigma_s^2+w_t\sigma_t^2)}{\Delta^2}$$

Therefore,

$$\sup_j \left\| \widehat{U}_j \widehat{U}_j^\top - U_j U_j^\top \right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr}\kappa\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2}}{\Delta} + \frac{pr(\kappa^2+\kappa_0^2)(w_s\sigma_s^2+w_t\sigma_t^2)}{\Delta^2} \right)$$

Using the same argument in the proof of Theorem 3.1, we have

$$\begin{aligned}
&\sup_j \left\| \widehat{V}_{j,t} \widehat{V}_{j,t}^\top - V_{j,t} V_{j,t}^\top \right\|_{\mathrm{F}} \\
&= O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}\kappa_0}{\lambda_{\min}} \Big(\sigma_t + \kappa^2\sqrt{r}\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2} \Big) + \frac{\sqrt{r_V} pr\kappa^2(\kappa^2+\kappa_0^2)(w_s\sigma_s^2+w_t\sigma_t^2)}{\lambda_{\min}^2} \right)
\end{aligned}$$

Since $\kappa_0, \kappa = O(1)$, $\Delta/\max(\sigma_s,\sigma_t) \gtrsim \sqrt{pr}$ and

$$(w_s\sigma_s^2+w_t\sigma_t^2) \le \sqrt{2}\max(\sigma_s,\sigma_t)\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2}$$

we have

$$\frac{pr(\kappa^2+\kappa_0^2)(w_s\sigma_s^2+w_t\sigma_t^2)}{\Delta^2} \lesssim \frac{\sqrt{pr}\kappa\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2}}{\Delta}$$

and thus

$$\sup_j \left\| \widehat{U}_j \widehat{U}_j^\top - U_j U_j^\top \right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr}\kappa\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2}}{\Delta} \right)$$

Similarly, it holds that

$$\frac{\sqrt{r_V} pr\kappa^2(\kappa^2+\kappa_0^2)(w_s\sigma_s^2+w_t\sigma_t^2)}{\lambda_{\min}^2} \lesssim \frac{\sqrt{pr_V}\kappa_0}{\lambda_{\min}} \Big(\kappa^2\sqrt{r}\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2} \Big)$$

which yields that

$$\sup_j \left\| \widehat{V}_{j,t} \widehat{V}_{j,t}^\top - V_{j,t} V_{j,t}^\top \right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}\kappa_0}{\lambda_{\min}} \Big(\sigma_t + \kappa^2\sqrt{r}\sqrt{w_s^2\sigma_s^2+w_t^2\sigma_t^2} \Big) \right)$$

Plugging in $w_s = \frac{\sigma_t^2}{\sigma_s^2+\sigma_t^2}$ and $w_t = \frac{\sigma_s^2}{\sigma_s^2+\sigma_t^2}$ leads to

$$\sup_j \left\| \widehat{U}_j \widehat{U}_j^\top - U_j U_j^\top \right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr}\kappa\sigma_s\sigma_t}{\Delta\sqrt{\sigma_s^2+\sigma_t^2}} \right)$$

and

$$\sup_j \left\| \widehat{V}_{j,t} \widehat{V}_{j,t}^\top - V_{j,t} V_{j,t}^\top \right\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\frac{\sqrt{pr_V}\kappa_0\sigma_t}{\lambda_{\min}} \Big(1 + \frac{\kappa^2\sqrt{r}\sigma_s}{\sqrt{\sigma_s^2+\sigma_t^2}} \Big) \right)$$

□

C.3 Proof of the Results for Distributed Inference

Proof for Theorem B.1. Following the proof of Theorem 2.1, for $j \in [J]$, $t \in \{1,2\}$, and $\ell \in [L]$, the estimator $\widehat{U}_{j,\ell}^{(t)}$ is composed of the first r_j eigenvectors of

$$(T_j + Z_{j,\ell})\widehat{U}_{-j}^{(t-1)}\widehat{U}_{-j}^{(t-1)\top}(T_j^\top + Z_{j,\ell}^\top) = T_j U_{-j} U_{-j}^\top T_j^\top + \mathfrak{E}_{j,\ell}^{(t)} = U_j \Lambda_j^2 U_j^\top + \mathfrak{E}_{j,\ell}^{(t)} \tag{40}$$

where $\mathfrak{E}_{j,\ell}^{(t)}$ is a remainder term defined by

$$\begin{aligned}
\mathfrak{E}_{j,\ell}^{(t)} &:= \zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top + \zeta_{j,\ell,2}^{(t)} + \zeta_{j,\ell,2}^{(t)\top} + \zeta_{j,\ell,3} + \zeta_{j,\ell,4}^{(t)} + \zeta_{j,\ell,5}^{(t)}, \\
\zeta_{j,\ell,1} &:= T_j U_{-j} U_{-j}^\top Z_{j,\ell}^\top, \\
\zeta_{j,\ell,2}^{(t)} &:= T_j \left[\widehat{U}_{-j}^{(t-1)}\widehat{U}_{-j}^{(t-1)\top} - U_{-j}U_{-j}^\top\right] Z_{j,\ell}^\top, \\
\zeta_{j,\ell,3} &:= Z_{j,\ell} U_{-j} U_{-j}^\top Z_{j,\ell}^\top, \\
\zeta_{j,\ell,4}^{(t)} &:= Z_{j,\ell} \left[\widehat{U}_{-j}^{(t-1)}\widehat{U}_{-j}^{(t-1)\top} - U_{-j}U_{-j}^\top\right] Z_{j,\ell}^\top, \\
\zeta_{j,\ell,5}^{(t)} &:= T_j \left[\widehat{U}_{-j}^{(t-1)}\widehat{U}_{-j}^{(t-1)\top} - U_{-j}U_{-j}^\top\right] T_j^\top
\end{aligned} \tag{41}$$

where

$$\begin{aligned}
U_{-j} &= U_1 \otimes U_2 \otimes \cdots \otimes U_{j-1} \otimes U_{j+1} \otimes \cdots \otimes U_J \\
\widehat{U}_{-j}^{(t)} &= \widehat{U}_1^{(t)} \otimes \widehat{U}_2^{(t)} \otimes \cdots \otimes \widehat{U}_{j-1}^{(t)} \otimes \widehat{U}_{j+1}^{(t)} \otimes \cdots \otimes \widehat{U}_J^{(t)}
\end{aligned}$$

For each $j \in [J]$, $\ell \in [L]$, define a "locally-good" event:

$$\begin{aligned}
&\widetilde{E}_{j,\ell}(C) \\
&:= \Bigg\{ \left\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} - U_j U_j^\top\right\|_2 \le C\sqrt{p}\sigma\lambda_{\min}^{-1}, \left\|\widehat{U}_{j,\ell}^{(1)}\widehat{U}_{j,\ell}^{(1)\top} - U_j U_j^\top\right\|_2 \le C\sqrt{p}\sigma\lambda_{\min}^{-1}, \|Z_{j,\ell}U_{-j}\|_2 \le C\sigma\sqrt{p}, \\
&\qquad \sup_{\substack{X_{j'} \in \mathbb{R}^{p_{j'} \times r_{j'}} \\ \|X_{j'}\|_2 \le 1, j' \in [J]\setminus\{j\}}} \|Z_{j,\ell}(X_1 \otimes X_2 \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J)\|_2 \le C\sigma\sqrt{pr} \Bigg\}
\end{aligned} \tag{42}$$

and a "globally-good" event

$$\widetilde{E}(C) := \bigcap_{\ell=1}^{L}\bigcap_{j=1}^{J} E_{j,\ell}(C)$$

By the proof of Lemma 1 and Theorem 2.1, we have $\mathbb{P}[\widetilde{E}(C_2)] \ge 1 - C_1 e^{-c_1 p}$ for some constants C_1, c_1, C_2. By Lemma 2, it holds that

$$\begin{aligned}
&\|\zeta_{j,\ell,1}\|_2 \le C_2'\kappa_0\lambda_{\min}\sigma\sqrt{p}, \quad \left\|\zeta_{j,\ell,2}^{(t)}\right\|_2 \le C_2'\kappa_0 p\sigma^2\sqrt{r}, \\
&\|\zeta_{j,\ell,3}\|_2 \le C_2' p\sigma^2, \quad \left\|\zeta_{j,\ell,4}^{(t)}\right\|_2 \le C_2' p^{3/2}\sqrt{r}\sigma^3\lambda_{\min}^{-1}, \quad \left\|\zeta_{j,\ell,5}^{(t)}\right\|_2 \le C_2'\kappa_0^2 p\sigma^2
\end{aligned} \tag{43}$$

for some absolute constant $C_2' > 0$, under the event $\widetilde{E}(C_2)$. Moreover, similar to (23), we have

$$\begin{aligned}
\widehat{U}_{j,\ell}^{(t)}\widehat{U}_{j,\ell}^{(t)\top} - U_j U_j^\top &= U_j\Lambda_j^{-2}U_j^\top \mathfrak{E}_{j,\ell}^{(t)} U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top \mathfrak{E}_{j,\ell}^{(t)} U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top \mathfrak{E}_{j,\ell}^{(t)} U_{j\perp}U_{j\perp}^\top \mathfrak{E}_{j,\ell}^{(t)} U_j\Lambda_j^{-2}U_j^\top - U_j\Lambda_j^{-2}U_j^\top \mathfrak{E}_{j,\ell}^{(t)} U_j\Lambda_j^{-2}U_j^\top \mathfrak{E}_{j,\ell}^{(t)} U_{j\perp}U_{j\perp}^\top \\
&\quad - U_{j\perp}U_{j\perp}^\top \mathfrak{E}_{j,\ell}^{(t)} U_j\Lambda_j^{-2}U_j^\top \mathfrak{E}_{j,\ell}^{(t)} U_j\Lambda_j^{-2}U_j^\top + \mathfrak{R}_{j,\ell}^{(t)}
\end{aligned} \tag{44}$$

where $\left\|\mathfrak{R}_{j,\ell}^{(t)}\right\|_2 \lesssim \kappa_0^3\sigma^3 p^{3/2}/\lambda_{\min}^3$. Then plugging (41) into (44) and using the fact that $U_{j\perp}^\top T_j = 0$, we obtain

$$\widehat{U}_{j,\ell}^{(t)}\widehat{U}_{j,\ell}^{(t)\top} - U_jU_j^\top = \mathfrak{S}_{j,\ell,1}^{(t)} + \mathfrak{S}_{j,\ell,2}^{(t)} + \mathfrak{S}_{j,\ell,3}^{(t)} \tag{45}$$

where for $j \in [J]$ and $t = 1, 2$,

$$\begin{aligned}
\mathfrak{S}_{j,\ell,1} &:= U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,1}U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,1}^\top U_j\Lambda_j^{-2}U_j^\top \\
&= U_j\Lambda_j^{-2}G_jU_{-j}^\top Z_{j,\ell}^\top U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top Z_{j,\ell}U_{-j}G_j^\top\Lambda_j^{-2}U_j^\top, \\
\mathfrak{S}_{j,\ell,2}^{(t)} &:= U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,2}^{(t)}U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,2}^{(t)\top}U_j\Lambda_j^{-2}U_j^\top \\
&\quad + U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,3}U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,3}U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,1}U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,1}^\top U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top(\zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top)U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,1}U_{j\perp}U_{j\perp}^\top \\
&\quad - U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,1}^\top U_j\Lambda_j^{-2}U_j^\top(\zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top)U_j\Lambda_j^{-2}U_j^\top
\end{aligned} \tag{46}$$

and

$$\begin{aligned}
\mathfrak{S}_{j,\ell,3}^{(t)} &:= U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,4}^{(t)}U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\zeta_{j,\ell,4}^{(t)}U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}^{(t)} - \zeta_{j,\ell,1} - \zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}^{(t)}U_{j\perp}U_{j\perp}^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}^{(t)}U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}^{(t)} - \zeta_{j,\ell,1} - \zeta_{j,\ell,1}^\top\right]U_{j\perp}U_{j\perp}^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}^{(t)} - \zeta_{j,\ell,1} - \zeta_{j,\ell,1}^\top\right]U_{j\perp}U_{j\perp}^\top\mathfrak{E}_{j,\ell}^{(t)}U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}^{(t)}U_{j\perp}U_{j\perp}^\top\left[\mathfrak{E}_{j,\ell}^{(t)} - \zeta_{j,\ell,1} - \zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_{j\perp}U_{j\perp}^\top\left[\mathfrak{E}_{j,\ell}^{(t)} - \zeta_{j,\ell,1} - \zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top\mathfrak{E}_{j,\ell}^{(t)}U_j\Lambda_j^{-2}U_j^\top \\
&\quad - U_{j\perp}U_{j\perp}^\top\mathfrak{E}_{j,\ell}^{(t)}U_j\Lambda_j^{-2}U_j^\top\left[\mathfrak{E}_{j,\ell}^{(t)} - \zeta_{j,\ell,1} - \zeta_{j,\ell,1}^\top\right]U_j\Lambda_j^{-2}U_j^\top + \mathfrak{R}_{j,\ell}^{(t)}
\end{aligned} \tag{47}$$

Then

$$\begin{aligned}
&\frac{1}{L}\sum_{\ell=1}^L \widehat{U}_{j,\ell}^{(t)}\widehat{U}_{j,\ell}^{(t)\top} - U_jU_j^\top \\
&= \overline{\mathfrak{S}_{j,1}} + \overline{\mathfrak{S}_{j,2}^{(t)}} + \overline{\mathfrak{S}_{j,3}^{(t)}} \\
&= U_j\Lambda_j^{-2}G_jU_{-j}^\top\overline{Z}_j^\top U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\Lambda_j^{-2}U_j^\top + \overline{\mathfrak{S}_{j,2}^{(t)}} + \overline{\mathfrak{S}_{j,3}^{(t)}}
\end{aligned} \tag{48}$$

where $\overline{Z}_j := (1/L)\sum_{\ell=1}^L Z_{j,\ell}$ and $\overline{\mathfrak{S}_{j,k}^{(t)}} := (1/L)\sum_{\ell=1}^L \mathfrak{S}_{j,\ell,k}^{(t)}$ for $k \in \{2, 3\}$ satisfies

$$\left\|\overline{\mathfrak{S}_{j,2}^{(t)}}\right\|_2 \lesssim \kappa_0^2 p r^{1/2}\sigma^2\lambda_{\min}^{-2}, \quad \left\|\overline{\mathfrak{S}_{j,3}^{(t)}}\right\|_2 \lesssim \kappa_0^3 p^{3/2} r^{1/2}\sigma^3\lambda_{\min}^{-3}$$

By (30), it holds that

$$\left\|\overline{\mathfrak{S}_{j,1}}\right\|_2 = \left\|U_j\Lambda_j^{-2}G_jU_{-j}^\top\overline{Z}_j^\top U_{j\perp}U_{j\perp}^\top + U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\Lambda_j^{-2}U_j^\top\right\|_2 \lesssim \lambda_{\min}^{-1}\sigma\sqrt{p/L}$$

with probability at least $1 - C_1e^{-c_1p}$.

Therefore, we obtain that

$$\frac{1}{L}\sum_{\ell=1}^{L}\widehat{U}_{j,\ell}^{(2)}\widehat{U}_{j,\ell}^{(2)\top}-U_jU_j^\top=\overline{\mathfrak{S}_{j,1}}+\overline{\mathfrak{S}_{j,2}^{(2)}}+\overline{\mathfrak{S}_{j,3}^{(2)}} \tag{49}$$

where

$$\left\|\overline{\mathfrak{S}_{j,1}}\right\|_2\lesssim\lambda_{\min}^{-1}\sigma\sqrt{p/L},\quad \left\|\overline{\mathfrak{S}_{j,2}^{(2)}}\right\|_2\lesssim\kappa_0^2pr^{1/2}\sigma^2\lambda_{\min}^{-2},\quad \left\|\overline{\mathfrak{S}_{j,3}^{(2)}}\right\|_2\lesssim\kappa_0^3p^{3/2}r^{1/2}\sigma^3\lambda_{\min}^{-3} \tag{50}$$

with probability at least $1-C_1'Le^{-c_1'p}$ for some absolute constant $C_1',c_1'>0$.

Since the columns of $\widehat{U}_j$ are the first r_j eigenvectors of $\frac{1}{L}\sum_{\ell=1}^{L}\widehat{U}_{j,\ell}^{(2)}\widehat{U}_{j,\ell}^{(2)\top}$, similar to (33),

$$\begin{aligned}\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top=&U_jU_j^\top\left(\overline{\mathfrak{S}_{j,1}}+\overline{\mathfrak{S}_{j,2}^{(2)}}\right)U_{j\perp}U_{j\perp}^\top+U_{j\perp}U_{j\perp}^\top\left(\overline{\mathfrak{S}_{j,1}}+\overline{\mathfrak{S}_{j,2}^{(2)}}\right)U_jU_j^\top\\&-U_jU_j^\top\overline{\mathfrak{S}_{j,1}}U_{j\perp}U_{j\perp}^\top\overline{\mathfrak{S}_{j,1}}U_jU_j^\top+\widetilde{\mathfrak{R}}_j\end{aligned} \tag{51}$$

where $\left\|\widetilde{\mathfrak{R}}_j\right\|_2\lesssim\kappa_0^3p^{3/2}r^{1/2}\sigma^3\lambda_{\min}^{-3}$. By (50), we obtain that

$$\left\|\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top\right\|_{\mathrm{F}}\le\sqrt{r}\left\|\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top\right\|_2=O_{\mathbb{P}}\left(\lambda_{\min}^{-1}\sigma\sqrt{pr/L}+\kappa_0^2pr\sigma^2\lambda_{\min}^{-2}\right)$$

Moreover,

$$\begin{aligned}&\left\|\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top\right\|_{\mathrm{F}}^2\\=&\left\langle\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top,\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top\right\rangle\\=&2\mathrm{tr}\left(U_j^\top\overline{\mathfrak{S}_{j,1}}U_{j\perp}U_{j\perp}^\top\overline{\mathfrak{S}_{j,1}}U_j\right)+4\mathrm{tr}\left(U_j^\top\overline{\mathfrak{S}_{j,2}^{(2)}}U_{j\perp}U_{j\perp}^\top\overline{\mathfrak{S}_{j,1}}U_j\right)+\mathfrak{Q}_j\\=&2\mathrm{tr}\left(\Lambda_j^{-4}G_jU_{-j}^\top\overline{Z}_j^\top U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\right)\\&+4\mathrm{tr}\left(\Lambda_j^{-4}U_j^\top\overline{\zeta_{j,2}^{(2)}}U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\right)\\&+4\mathrm{tr}\left(\Lambda_j^{-4}U_j^\top\overline{\zeta_{j,3}}U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\right)\\&-4\mathrm{tr}\left\{\Lambda_j^{-4}U_j^\top\left[\frac{1}{L}\sum_{\ell=1}^{L}\left(\zeta_{j,\ell,1}+\zeta_{j,\ell,1}^\top\right)U_j\Lambda_j^{-2}U_j^\top\zeta_{j,\ell,1}\right]U_{j\perp}U_{j\perp}^\top\overline{Z}_jU_{-j}G_j^\top\right\}+\mathfrak{Q}_j\end{aligned} \tag{52}$$

where $\overline{\zeta_{j,k}^{(t)}}:=(1/L)\sum_{\ell=1}^{L}\zeta_{j,\ell,k}^{(t)}$ and $\mathfrak{Q}_j$ is a remainder term. By the inequality that $|\langle A,B\rangle|\le r\|A\|_2\|B\|_2$ for all rank-r matrices A and B, we bound the remainder term by $|\mathfrak{Q}_j|\lesssim\kappa_0^4p^2r^2\sigma^4\lambda_{\min}^{-4}$.

Now we provide bounds for the second, third and fourth terms in (52). By (45),

$$U_j^\top\widehat{U}_{j,\ell}^{(1)}\widehat{U}_{j,\ell}^{(1)\top}=U_j^\top+\Lambda_j^{-2}G_jU_{-j}^\top Z_{j,\ell}^\top U_{j\perp}U_{j\perp}^\top+U_j^\top\left(\overline{\mathfrak{S}_{j,2}^{(1)}}+\overline{\mathfrak{S}_{j,3}^{(1)}}\right)$$

Hence,

$$\begin{aligned}
&\zeta^{(2)}_{j,\ell,2} \\
&= T_j \left[\widehat{U}^{(1)}_{-j}\widehat{U}^{(1)\top}_{-j} - U_{-j}U^\top_{-j}\right] Z^\top_{j,\ell} \\
&= U_j G_j U^\top_{-j} \left[\widehat{U}^{(1)}_{-j}\widehat{U}^{(1)\top}_{-j} - U_{-j}U^\top_{-j}\right] Z^\top_{j,\ell} \\
&= U_j G_j \left[\bigotimes_{j'\neq j}\left(U^\top_{j'}\widehat{U}^{(1)}_{j',\ell}\widehat{U}^{(1)\top}_{j',\ell}\right) - U^\top_{-j}\right] Z^\top_{j,\ell} \\
&= U_j G_j \left[\sum_{j'\neq j} U_1^\top \otimes \cdots U^\top_{j'-1} \otimes \left(\Lambda^{-2}_{j'} G_{j'} U_{-j'} Z^\top_{j',\ell} U_{j'\perp} U^\top_{j'\perp}\right) \otimes U^\top_{j'+1}\cdots U^\top_{j-1} \otimes U^\top_{j+1}\cdots \otimes U^\top_J\right] Z^\top_{j,\ell} \\
&\quad + \mathfrak{R}_{\zeta^{(2)}_{j,\ell,2}}
\end{aligned} \tag{53}$$

where $\left\|\mathfrak{R}_{\zeta^{(2)}_{j,\ell,2}}\right\|_2 \leq C_4\kappa_0^3 p^{3/2} r^{1/2}\sigma^3\lambda^{-1}_{\min}$ for some $C_4 > 0$. As a result, we obtain

$$\operatorname{tr}\left(\Lambda_j^{-4} U_j^\top \overline{\zeta^{(2)}_{j,2}} U_{j\perp} U^\top_{j\perp} \overline{Z_j} U_{-j} G_j^\top\right) = \sum_{j'\neq j} \mathfrak{M}_{j'} + \mathfrak{R}_{\mathfrak{M}} \tag{54}$$

where

$$\begin{aligned}
\mathfrak{M}_{j'} := \operatorname{tr}\Bigg\{ &\Lambda_j^{-4} G_j \\
&\cdot \left[\frac{1}{L}\sum_{\ell=1}^L \left(U_1^\top \otimes \cdots \otimes U^\top_{j'-1} \otimes \left(\Lambda^{-2}_{j'} G_{j'} U_{-j'} Z^\top_{j',\ell} U_{j'\perp} U^\top_{j'\perp}\right) \otimes U^\top_{j'+1}\cdots \otimes U^\top_{j-1} \otimes U^\top_{j+1} \otimes \cdots \otimes U^\top_J\right) Z^\top_{j,\ell}\right] \cdot \\
&\quad U_{j\perp} U^\top_{j\perp} \overline{Z_j} U_{-j} G_j^\top \Bigg\}
\end{aligned} \tag{55}$$

and $|\mathfrak{R}_{\mathfrak{M}}| \lesssim \kappa_0^4 p^2 r^2 \sigma^4 \lambda^{-4}_{\min}$. Define

$$W_{j',\ell,a} := U^\top_{j'\perp} Z_{j',\ell} U_{-j'}$$

$$W_{j',\ell,b} := U^\top_{j\perp} Z_{j,\ell}(U_1 \otimes \cdots \otimes U_{j'-1} \otimes U_{j'\perp} \otimes U_{j'+1} \otimes \cdots \otimes U_{j-1} \otimes U_{j+1} \otimes \cdots \otimes U_J)$$

$$W_{j,\ell,c} := U^\top_{j\perp} Z_{j,\ell} U_{-j}$$

then $\mathfrak{M}_{j'}$ can be simplified as

$$\mathfrak{M}_{j'} = \operatorname{tr}\left[\Lambda_j^{-4} G_j \frac{1}{L}\sum_{\ell=1}^L \left(I_{r_1} \otimes \cdots I_{r_{j'-1}} \otimes \left(\Lambda^{-2}_{j'} G_{j'} W^\top_{j',\ell,a}\right) \otimes I_{r_{j'+1}} \cdots \otimes I_{r_J}\right) W^\top_{j',\ell,b} \overline{W_{j,c}} G_j^\top\right]$$

where $\overline{W_{j,c}} = (1/L)\sum_{\ell=1}^L W_{j,\ell,c}$.

Let vec denote the vectorization of a matrix. By assumption, $\operatorname{vec}(Z_{j',\ell}) \sim \mathcal{N}(0, \sigma^2 I)$. Using the identity that $\operatorname{vec}(ABC) = (C^\top \otimes A)\operatorname{vec}(B)$ for all matrices A, B, C, we obtain that $\operatorname{vec}(W_{j',\ell,a}) \sim \mathcal{N}(0, \sigma^2 I)$, i.e., the entries of $W_{j',\ell,a}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Similarly, $\operatorname{vec}(W_{j',\ell,b}) \sim \mathcal{N}(0, \sigma^2 I)$ and

$\mathrm{vec}(W_{j,\ell,c}) \sim \mathcal{N}(0,\sigma^2 I)$. Furthermore, since $U_{j'\perp}^\top U_{j'} = 0$, $\mathbb{E}[\mathrm{vec}(W_{j',\ell,b})\mathrm{vec}(W_{j,\ell,c})^\top] = 0$, which implies that $W_{j',\ell,b}$ and $W_{j,\ell,c}$ are independent. Therefore, conditional on $\{W_{j',\ell,a}, W_{j,\ell,c}\}_{\ell=1}^L$,

$$\begin{aligned}&\mathfrak{M}_{j'} \mid \{W_{j',\ell,a}, W_{j,\ell,c}\}_{\ell=1}^L \\ &\sim \mathcal{N}\left(0, \frac{\sigma^2}{L^2}\sum_{\ell=1}^L \left\|\overline{W_{j,c}} G_j^\top \Lambda_j^{-4} G_j \left(I_{r_1}\otimes\cdots I_{r_{j'-1}}\otimes\left(\Lambda_{j'}^{-2}G_{j'}W_{j',\ell,a}^\top\right)\otimes I_{r_{j'+1}}\cdots\otimes I_{r_J}\right)\right\|_F^2\right)\end{aligned} \tag{56}$$

By the proof of Theorem 2.1, with probability at least $1 - C_1 e^{-c_1 p}$, $\left\|W_{j',\ell,a}\right\|_2 \le C_2\sigma\sqrt{p}$ and $\left\|\overline{W_{j,c}}\right\|_2 \le C_2\sigma\sqrt{p/L}$. Note that

$$\mathrm{rank}\left[\overline{W_{j,c}} G_j^\top \Lambda_j^{-4} G_j \left(I_{r_1}\otimes\cdots I_{r_{j'-1}}\otimes\left(\Lambda_{j'}^{-2}G_{j'}W_{j',\ell,a}^\top\right)\otimes I_{r_{j'+1}}\cdots\otimes I_{r_J}\right)\right] \le \mathrm{rank}(G_j) = r_j$$

Then, using that $\left\|\Lambda_j^{-1}G_j\right\|_2 = 1$, we obtain,

$$\left\|\overline{W_{j,c}} G_j^\top \Lambda_j^{-4} G_j \left(I_{r_1}\otimes\cdots I_{r_{j'-1}}\otimes\left(\Lambda_{j'}^{-2}G_{j'}W_{j',\ell,a}^\top\right)\otimes I_{r_{j'+1}}\cdots\otimes I_{r_J}\right)\right\|_F \le C_2^2 p r^{1/2} L^{-1/2}\sigma^2\lambda_{\min}^{-3}$$

Hence,

$$\begin{aligned}\mathrm{sd}(\mathfrak{M}_{j'}) &:= \sqrt{\frac{\sigma^2}{L^2}\sum_{\ell=1}^L \left\|\overline{W_{j,c}} G_j^\top \Lambda_j^{-4} G_j \left(I_{r_1}\otimes\cdots I_{r_{j'-1}}\otimes\left(\Lambda_{j'}^{-2}G_{j'}W_{j',\ell,a}^\top\right)\otimes I_{r_{j'+1}}\cdots\otimes I_{r_J}\right)\right\|_F^2} \\ &\lesssim L^{-1} p r^{1/2}\sigma^3\lambda_{\min}^{-3}\end{aligned}$$

By (56), for any $\gamma > 0$,

$$\mathbb{P}\left(\left|\mathfrak{M}_{j'}\right| \ge \mathrm{sd}(\mathfrak{M}_{j'})\sqrt{2\gamma\log p} \quad \middle| \quad \{W_{j',\ell,a}, W_{j,\ell,c}\}_{\ell=1}^L\right) \le 2p^{-\gamma}$$

Therefore, with probability at least $1 - C_1 L e^{-c_1 p} - 2p^{-\gamma}$,

$$\left|\mathfrak{M}_{j'}\right| \lesssim \frac{p\sigma^3\sqrt{\gamma r\log p}}{\lambda_{\min}^3 L}$$

In conclusion, the second term in (52) can be bounded by

$$\left|4\mathrm{tr}\left(\Lambda_j^{-4}U_j^\top\overline{\zeta_{j,2}^{(2)}}U_{j\perp}U_{j\perp}^\top\overline{Z_j}U_{-j}G_j^\top\right)\right| \lesssim \frac{p\sigma^3\sqrt{\gamma r\log p}}{\lambda_{\min}^3 L} + \kappa_0^4 p^2 r^2\sigma^4\lambda_{\min}^{-4}, \tag{57}$$

with probability at least $1 - C_1 J L e^{-c_1 p} - 2Jp^{-\gamma}$.

Next, we deal with the third term in (52), i.e.,

$$\mathrm{tr}\left(\Lambda_j^{-4}U_j^\top\overline{\zeta_{j,3}}U_{j\perp}U_{j\perp}^\top\overline{Z_j}U_{-j}G_j^\top\right) = \frac{1}{L}\sum_{\ell=1}^L \mathrm{tr}\left[\Lambda_j^{-4}W_{j,\ell,d}W_{j,\ell,c}^\top\overline{W_{j,c}}G_j^\top\right] := \widetilde{\mathfrak{M}_0}$$

where $W_{j,\ell,d} := U_j^\top Z_{j,\ell}U_{-j}$. Since $\mathrm{vec}([W_{j,\ell,d}W_{j,\ell,c}]) = \mathrm{vec}([U_jU_{j\perp}]^\top Z_{j,\ell}U_{-j}) \sim \mathcal{N}(0,\sigma^2 I)$, we have

$$\widetilde{\mathfrak{M}_0} \mid \{W_{j,\ell,c}\}_{\ell=1}^L \sim \mathcal{N}\left(0, \frac{\sigma^2}{L^2}\sum_{\ell=1}^L \left\|\Lambda_j^{-4}G_j\overline{W_{j,c}}^\top W_{j,\ell,c}\right\|_{\mathrm{F}}^2\right)$$

Using that $\|W_{j,\ell,c}\|_2 \le C_2\sigma\sqrt{p}$ and $\left\|\overline{W_{j,c}}\right\|_2 \le C_2\sigma\sqrt{p/L}$ with probability at least $1 - C_1 e^{-c_1 p}$, we have $\mathrm{sd}(\widetilde{\mathfrak{M}}_0) \lesssim \sigma^3 L^{-1} r^{1/2} p \lambda_{\min}^{-3}$. Therefore, with probability at least $1 - C_1 L e^{-c_1 p} - 2p^{-\gamma}$, we obtain that

$$\left|\widetilde{\mathfrak{M}}_0\right| \lesssim \frac{p\sigma^3\sqrt{\gamma r \log p}}{\lambda_{\min}^3 L} \tag{58}$$

For the fourth term in (52), note that

$$\begin{aligned}
&\mathrm{tr}\left\{\Lambda_j^{-4} U_j^\top \left[\frac{1}{L}\sum_{\ell=1}^{L}\left(\zeta_{j,\ell,1} + \zeta_{j,\ell,1}^\top\right) U_j \Lambda_j^{-2} U_j^\top \zeta_{j,\ell,1}\right] U_{j\perp} U_{j\perp}^\top \overline{Z_j} U_{-j} G_j^\top\right\} \\
&= \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{tr}\left[\Lambda_j^{-4} G_j W_{j,\ell,d}^\top \Lambda_j^{-2} G_j W_{j,\ell,c}^\top \overline{W_{j,c}} G_j^\top\right] + \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{tr}\left[\Lambda_j^{-4} W_{j,\ell,d} G_j^\top \Lambda_j^{-2} G_j W_{j,\ell,c}^\top \overline{W_{j,c}} G_j^\top\right] \\
&=: \widetilde{\mathfrak{M}}_1 + \widetilde{\mathfrak{M}}_2
\end{aligned} \tag{59}$$

Repeating the analysis for $\widetilde{\mathfrak{M}}_0$ yields the same result that

$$\left|\widetilde{\mathfrak{M}}_1\right| \lesssim \frac{p\sigma^3\sqrt{\gamma r \log p}}{\lambda_{\min}^3 L}, \quad \left|\widetilde{\mathfrak{M}}_2\right| \lesssim \frac{p\sigma^3\sqrt{\gamma r \log p}}{\lambda_{\min}^3 L} \tag{60}$$

Combining (52), (57), (58) and (60) leads to that

$$\begin{aligned}
&\left|\left\|\widehat{U}_j\widehat{U}_j^\top - U_j U_j^\top\right\|_{\mathrm{F}}^2 - 2\mathrm{tr}\left(\Lambda_j^{-4} G_j U_{-j}^\top \overline{Z}_j^\top U_{j\perp} U_{j\perp}^\top \overline{Z}_j U_{-j} G_j^\top\right)\right| \\
&= O_{\mathbb{P}}\left(\frac{p\sigma^3\sqrt{r\log p}}{\lambda_{\min}^3 L} + \frac{\kappa_0^4 p^2 r^2 \sigma^4}{\lambda_{\min}^4}\right)
\end{aligned} \tag{61}$$

Now we focus on the first term. By the proof of the final step of Theorem 1 in [XZZ22], it holds that

$$\begin{aligned}
&2\mathrm{tr}\left(\Lambda_j^{-4} G_j U_{-j}^\top \overline{Z}_j^\top U_{j\perp} U_{j\perp}^\top \overline{Z}_j U_{-j} G_j^\top\right) \\
&= 2\left\|\Lambda_j^{-2} G_j U_{-j}^\top \overline{Z}_j^\top U_{j\perp}\right\|_{\mathrm{F}}^2 \\
&\overset{d}{=} \frac{2\sigma^2}{L}\sum_{i=1}^{p_j - r_j}\left\|\Lambda_j^{-1}\boldsymbol{z}_i\right\|_2^2
\end{aligned}$$

where $\boldsymbol{z}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{r_j})$. Since $\mathbb{E}\left[\|\Lambda_1^{-1}\boldsymbol{z}_i\|_2^2\right] = \left\|\Lambda_j^{-1}\right\|_{\mathrm{F}}^2$ and $\mathrm{Var}\left[\left\|\Lambda_j^{-1}\boldsymbol{z}_i\right\|_2^2\right] = 2\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}}^2$, by Central Limit Theorem,

$$\frac{2\left\|\Lambda_j^{-2} G_j U_{-j}^\top \overline{Z}_j^\top U_{j\perp}\right\|_{\mathrm{F}}^2 - 2\sigma^2 L^{-1}(p_j - r_j)\left\|\Lambda_j^{-1}\right\|_{\mathrm{F}}^2}{\sqrt{8(p_j - r_j)}\sigma^2 L^{-1}\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}}} \xrightarrow{d} \mathcal{N}(0,1)$$

Since $\sqrt{8(p_j - r_j)}\sigma^2 L^{-1}\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}} \gtrsim \sqrt{p_j r_j} L^{-1}\sigma^2\kappa_0^{-2}\lambda_{\min}^{-2}$ and $r_j/p_j = o(1)$, by (61), it holds that

$$\frac{\left\|\widehat{U}_j\widehat{U}_j^\top - U_j U_j^\top\right\|_{\mathrm{F}}^2 - 2\sigma^2 L^{-1}(p_j - r_j)\left\|\Lambda_j^{-1}\right\|_{\mathrm{F}}^2}{\sqrt{8p_j}\sigma^2 L^{-1}\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}}} \xrightarrow{d} \mathcal{N}(0,1)$$

if

$$\left(\frac{p\sigma^3\sqrt{r\log p}}{\lambda_{\min}^3 L}+\frac{\kappa_0^4 p^2 r^2\sigma^4}{\lambda_{\min}^4}\right)/\frac{\sqrt{p_j r_j}\sigma^2}{L\kappa_0^2\lambda_{\min}^2}=o(1)$$

or

$$\frac{Lr^{3/2}\kappa_0^6 p^{3/2}}{(\lambda_{\min}/\sigma)^2}+\frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}/\sigma}=o(1)$$

Furthermore, if $r^3/p=o(1)$, we have $r_j\left\|\Lambda_j^{-1}\right\|_{\mathrm{F}}^2/\sqrt{p_j}\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}}=o(1)$, then

$$\frac{\left\|\widehat{U}_j\widehat{U}_j^\top-U_jU_j^\top\right\|_{\mathrm{F}}^2-2\sigma^2L^{-1}p_j\left\|\Lambda_j^{-1}\right\|_{\mathrm{F}}^2}{\sqrt{8p_j}\sigma^2L^{-1}\left\|\Lambda_j^{-2}\right\|_{\mathrm{F}}}\xrightarrow{d}\mathcal{N}(0,1)$$

□

C.4 Proof of Technical Lemmas

Proof of Lemma 1. We first show the high-probability bound for a fix (j,ℓ). It is assumed in Theorem 2.1 that there exist C_1,c_1,C_2 such that

$$\mathbb{P}\left[\left\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top}-U_jU_j^\top\right\|_2\le C_2\sqrt{p}\sigma\lambda_{\min}^{-1}\right]\ge 1-C_1e^{-c_1p} \tag{62}$$

Since $Z_{j,\ell}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries and $U_{-j}^\top U_{-j}=I_{r_{-j}}$, the matrix $Z_{j,\ell}U_{-j}$ also has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. Then we need the following lemma:

Lemma 3. *[Theorem 5.39 in [Ver10]] Let $M\in\mathbb{R}^{p_1\times p_2}$ whose rows $M_1,\dots,M_{p_1}$ are independent sub-Gaussian random vectors with $\mathbb{E}[M_iM_i^\top]=I_{p_2}$. Then for every $t\ge 0$, there exist constants $c,C>0$ such that*

$$\mathbb{P}\left[\|M\|_2\le\sqrt{p_1}+C\sqrt{p_2}+t\right]\ge 1-2e^{-ct^2}$$

for any $t\ge 0$.

By Lemma 3, for any $t>0$, there exist c,C such that

$$\left\|\frac{1}{\sigma}Z_{j,\ell}U_{-j}\right\|_2\le\sqrt{p_j}+C\sqrt{r_{-j}}+t$$

with probability at least $1-2e^{-ct^2}$. Using the assumptions $p\asymp p_j$ and $p\gtrsim r^{J-1}\ge r_{-j}$ and choosing $t=\sqrt{p}$, we obtain that

$$\mathbb{P}\left[\|Z_{j,\ell}U_{-j}\|_2\le C_2\sigma\sqrt{p}\right]\ge 1-2e^{-c_1p} \tag{63}$$

for some $c_1,C_2>0$.

To show the third inequality in (19), we first define $\mathcal{B}^{p\times r}(X_0,\varepsilon):=\{X\mid X\in\mathbb{R}^{p\times r},\|X-X_0\|_2\le\varepsilon\}$. By Lemma 7 in [ZX18], for any $\varepsilon>0$, there exist $\mathcal{C}^{p,r}:=\{\overline{X}_1,\dots,\overline{X}_N\}$ such that $\left\|\overline{X}_i\right\|_2\le 1$ and $\mathcal{B}^{p\times r}(\mathbf{0},1)\subset\bigcup_{i=1}^N\mathcal{B}^{p\times r}(\overline{X}_i,\varepsilon)$, with $N\le(1+2/\varepsilon)^{pr}$. In particular, let $\varepsilon=1/(2J)$, and we

have that, for any $(X_1, \ldots, X_J)$ satisfying $X_j \in \mathcal{B}^{p_j \times r_j}(\mathbf{0}, 1)$, there exists $\big(\widetilde{X}_1, \ldots, \widetilde{X}_J\big) \in \mathcal{C}^{p_1,r_1} \times \mathcal{C}^{p_2,r_2} \times \cdots \times \mathcal{C}^{p_J,r_J}$ such that $\left\|X_j - \widetilde{X}_j\right\|_2 \le 1/(2J)$ for all j. Since

$$
\begin{aligned}
&\|Z_{j,\ell}(X_1 \otimes X_2 \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J)\|_2 - \left\|Z_{j,\ell}(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J)\right\|_2 \\
&\le \sum_{j' \in [J]\setminus\{j\}} \left\|Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes (X_{j'} - \widetilde{X}_{j'}) \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J\big)\right\|_2 \\
&\le \sum_{j' \in [J]\setminus\{j\}} \left\|X_{j'} - \widetilde{X}_{j'}\right\|_2 \sup_{\substack{X_{j'} \in \mathbb{R}^{p_{j'} \times r_{j'}} \\ \|X_{j'}\|_2 \le 1, j' \in [J]\setminus\{j\}}} \left\|Z_{j,\ell}\big(X_1 \otimes X_2 \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J\big)\right\|_2 \\
&\le \frac{1}{2} \sup_{\substack{X_{j'} \in \mathbb{R}^{p_{j'} \times r_{j'}} \\ \|X_{j'}\|_2 \le 1, j' \in [J]\setminus\{j\}}} \left\|Z_{j,\ell}\big(X_1 \otimes X_2 \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J\big)\right\|_2
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\sup_{\substack{X_{j'} \in \mathbb{R}^{p_{j'} \times r_{j'}} \\ \|X_{j'}\|_2 \le 1, j' \in [J]\setminus\{j\}}} \left\|Z_{j,\ell}\big(X_1 \otimes X_2 \otimes \cdots \otimes X_{j-1} \otimes X_{j+1} \otimes \cdots \otimes X_J\big)\right\|_2 \\
&\le 2 \sup_{\left(\widetilde{X}_1, \ldots, \widetilde{X}_J\right) \in \mathcal{C}^{p_1,r_1} \times \cdots \times \mathcal{C}^{p_J,r_J}} \left\|Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J\big)\right\|_2
\end{aligned} \tag{64}
$$

Therefore, it suffices to show that there exist C_1, c_1, C_2, s.t.

$$
\sup_{\left(\widetilde{X}_1, \ldots, \widetilde{X}_J\right) \in \mathcal{C}^{p_1,r_1} \times \cdots \times \mathcal{C}^{p_J,r_J}} \left\|Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J\big)\right\|_2 \le C_2 \sigma \sqrt{pr}
$$

with probability at least $1 - C_1 e^{-c_1 p}$.

For a fixed tuple $\big(\widetilde{X}_1, \ldots, \widetilde{X}_J\big)$, let $\Psi_{j,\ell} = Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J\big)$. The rows of $\Psi_{j,\ell}$ are independent normal vectors $\mathcal{N}(0, \Sigma_{-j})$, where $\Sigma_{-j} = \sigma^2\big(\widetilde{X}_1^\top \widetilde{X}_1 \otimes \widetilde{X}_2^\top \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1}^\top \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1}^\top \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J^\top \widetilde{X}_J\big)$, and hence $\Psi_{j,\ell}\Sigma_{-j}^{-1/2}$ has i.i.d. standard normal entries. Applying Lemma 3 to $\Psi_{j,\ell}\Sigma_{-j}^{-1/2}$ and using that $\|\Psi_{j,\ell}\|_2 \le \left\|\Psi_{j,\ell}\Sigma_{-j}^{-1/2}\right\|_2 \left\|\Sigma_{-j}^{-1/2}\right\|_2 \le \sigma \left\|\Psi_{j,\ell}\Sigma_{-j}^{-1/2}\right\|_2$ leads to

$$
\left\|Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J\big)\right\|_2 \le \sigma(\sqrt{p_j} + C\sqrt{r_{-j}} + t)
$$

with probability at least $1 - 2e^{-ct^2}$, for some absolute constants $c, C > 0$. Using the assumptions $p \asymp p_j$ and $p \gtrsim r^{J-1} \ge r_{-j}$ and letting $t = \gamma'\sqrt{pr}$, we obtain

$$
\left\|Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J\big)\right\|_2 \le C'\sigma\sqrt{pr}
$$

with probability at least $1 - 2e^{-c\gamma'^2 pr}$, for any $\gamma' > 0$ and some $C' > 0$ depending on γ'. This further leads to that

$$
\sup_{\left(\widetilde{X}_1, \ldots, \widetilde{X}_J\right) \in \mathcal{C}^{p_1,r_1} \times \cdots \times \mathcal{C}^{p_J,r_J}} \left\|Z_{j,\ell}\big(\widetilde{X}_1 \otimes \widetilde{X}_2 \otimes \cdots \otimes \widetilde{X}_{j-1} \otimes \widetilde{X}_{j+1} \otimes \cdots \otimes \widetilde{X}_J\big)\right\|_2 \le C'\sigma\sqrt{pr}
$$

with probability at least $1-2(4J+1)^{Jpr}e^{-c\gamma'^2 pr}$. By choosing γ' sufficiently large and combining the above inequality with (64), we finally obtain that there exist $C_1, c_1, C_2 > 0$, such that

$$\sup_{\substack{X_{j'}\in\mathbb{R}^{p_{j'}\times r_{j'}}\\ \|X_{j'}\|_2\le 1, j'\in[J]\setminus\{j\}}} \left\|Z_{j,\ell}\big(X_1\otimes X_2\otimes\cdots\otimes X_{j-1}\otimes X_{j+1}\otimes\cdots\otimes X_J\big)\right\|_2 \le C_2\sigma\sqrt{pr} \tag{65}$$

with probability at least $1-C_1e^{-c_1pr}$. Combining (62), (63), and (65) leads to $\mathbb{P}[E_{j,\ell}(C_2)] \ge 1-C_1e^{-c_1p}$, implying that $\mathbb{P}[E(C_2)] \ge 1-C_1Le^{-c_1p}$, for some constants $C_1, c_1, C_2 > 0$. □

Proof of Lemma 2. Under event $E(C_2)$, it holds that $\sup_{j,\ell}\|Z_{j,\ell}U_{-j}\|_2 \le C_2\sigma\sqrt{p}$. Combining with $\|T_j\|_2 = \|G_j\|_2 \le \kappa_0\lambda_{\min}$, it is straightforward to show that $\|\zeta_{j,\ell,1}\|_2 \le C_2\kappa_0\lambda_{\min}\sigma\sqrt{p}$ and $\|\zeta_{j,\ell,3}\|_2 \le \|Z_{j,\ell}U_{-j}\|_2^2 \le C_2^2\sigma^2 p$.

We have the following decomposition

$$\begin{aligned}
&\widehat{U}^{(0)}_{-j,\ell}\widehat{U}^{(0)\top}_{-j,\ell} - U_{-j}U_{-j}^\top\\
&= \sum_{j'\neq j}\Bigg[\left(U_1U_1^\top\right)\otimes\cdots\otimes\left(U_{j'-1}U_{j'-1}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell} - U_{j'}U_{j'}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j'+1,\ell}\widehat{U}^{(0)\top}_{j'+1,\ell}\right)\\
&\qquad\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{j-1,\ell}\widehat{U}^{(0)\top}_{j-1,\ell}\right)\otimes\left(\widehat{U}^{(0)}_{j+1,\ell}\widehat{U}^{(0)\top}_{j+1,\ell}\right)\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{J,\ell}\widehat{U}^{(0)\top}_{J,\ell}\right)\Bigg]\\
&= \sum_{j'\neq j}\Bigg[\left(U_1U_1^\top\right)\otimes\cdots\otimes\left(U_{j'-1}U_{j'-1}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell} - \widehat{U}^{(0)}_{j',\ell}B_{j',\ell}A_{j',\ell}^\top U_{j'}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j'+1,\ell}\widehat{U}^{(0)\top}_{j'+1,\ell}\right)\\
&\qquad\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{j-1,\ell}\widehat{U}^{(0)\top}_{j-1,\ell}\right)\otimes\left(\widehat{U}^{(0)}_{j+1,\ell}\widehat{U}^{(0)\top}_{j+1,\ell}\right)\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{J,\ell}\widehat{U}^{(0)\top}_{J,\ell}\right)\\
&\qquad+\left(U_1U_1^\top\right)\otimes\cdots\otimes\left(U_{j'-1}U_{j'-1}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j',\ell}B_{j',\ell}A_{j',\ell}^\top U_{j'}^\top - U_{j'}U_{j'}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j'+1,\ell}\widehat{U}^{(0)\top}_{j'+1,\ell}\right)\\
&\qquad\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{j-1,\ell}\widehat{U}^{(0)\top}_{j-1,\ell}\right)\otimes\left(\widehat{U}^{(0)}_{j+1,\ell}\widehat{U}^{(0)\top}_{j+1,\ell}\right)\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{J,\ell}\widehat{U}^{(0)\top}_{J,\ell}\right)\Bigg]\\
&= \sum_{j'\neq j}\Bigg[\left(U_1\otimes\cdots\otimes U_{j'-1}\otimes\widehat{U}^{(0)}_{j',\ell}\otimes\widehat{U}^{(0)}_{j'+1}\otimes\cdots\otimes\widehat{U}^{(0)}_{J}\right)\\
&\quad\cdot\left(U_1^\top\otimes\cdots\otimes U_{j'-1}^\top\otimes\left(\widehat{U}^{(0)\top}_{j',\ell} - B_{j',\ell}A_{j',\ell}^\top U_{j'}^\top\right)\otimes\widehat{U}^{(0)\top}_{j'+1}\otimes\cdots\otimes\widehat{U}^{(0)\top}_{J}\right)\\
&\qquad+\left(U_1\otimes\cdots\otimes U_{j'-1}\otimes\left(\widehat{U}^{(0)}_{j',\ell}B_{j',\ell}A_{j',\ell}^\top - U_{j'}\right)\otimes\widehat{U}^{(0)}_{j'+1}\otimes\cdots\otimes\widehat{U}^{(0)}_{J}\right)\\
&\quad\cdot\left(U_1^\top\otimes\cdots\otimes U_{j'-1}^\top\otimes U_{j'}^\top\otimes\widehat{U}^{(0)\top}_{j'+1}\otimes\cdots\otimes\widehat{U}^{(0)\top}_{J}\right)
\end{aligned} \tag{66}$$

where $A_{j,\ell}\in\mathbb{O}^{r_j\times r_j}$ and $B_{j,\ell}\in\mathbb{O}^{r_j\times r_j}$ are defined by an SVD for $U_j^\top\widehat{U}^{(0)}_{j,\ell}$, that is, $U_j^\top\widehat{U}^{(0)}_{j,\ell} =$

$A_{j,\ell}S_{j,\ell}B_{j,\ell}^\top$. Note that

$$\begin{aligned}
&\left\|\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top - U_{j'}\right\|_2 \\
&= \left\|\begin{pmatrix} U_{j'}^\top \\ U_{j'\perp}^\top \end{pmatrix}\left(\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top - U_{j'}\right)\right\|_2 \\
&= \left\|\begin{pmatrix} U_{j'}^\top\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top - I_{r_{j'}} \\ U_{j'\perp}^\top\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top \end{pmatrix}\right\|_2 \\
&\le \sqrt{\left\|U_{j'}^\top\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top - I_{r_{j'}}\right\|_2^2 + \left\|U_{j'\perp}^\top\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top\right\|_2^2} \\
&\le \sqrt{\left[1-\lambda_{\min}\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right)\right]^2 + 1 - \lambda_{\min}^2\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right)} \\
&\le \sqrt{2}\sqrt{1-\lambda_{\min}^2\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right)} \le \sqrt{2}\left\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} - U_jU_j^\top\right\|_2
\end{aligned}$$

where we sequentially use the following properties:

(1) For all block matrix $[A\ B]$, we have

$$\|[A\ B]\|_2 = \sup_{\|\boldsymbol{v}\|_2=1}\|[A\ B]\boldsymbol{v}\|_2 = \sup_{\|\boldsymbol{v}_1\|_2^2+\|\boldsymbol{v}_2\|_2^2=1}\|A\boldsymbol{v}_1\|_2 + \|B\boldsymbol{v}_2\|_2 \le \sqrt{\|A\|_2^2+\|B\|_2^2}$$

(2) For all $U, \widehat{U} \in \mathbb{O}^{p\times r}$, we have

$$\left\|U_\perp^\top\widehat{U}\right\|_2 = \sup_{\|\boldsymbol{v}\|_2=1}\left\|U_\perp^\top\widehat{U}\boldsymbol{v}\right\|_2 = \sup_{\|\boldsymbol{v}\|_2=1}\sqrt{\left\|\widehat{U}\boldsymbol{v}\right\|_2^2 - \left\|U^\top\widehat{U}\boldsymbol{v}\right\|_2^2} = \sqrt{1-\lambda_{\min}^2\left(U^\top\widehat{U}\right)}$$

(3) As $\left\|U_j^\top\widehat{U}_{j,\ell}^{(0)}\right\|_2 \le 1$, we have that $\lambda_{\min}\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right) \le 1$, and hence

$$\sqrt{\left[1-\lambda_{\min}\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right)\right]^2 + 1 - \lambda_{\min}^2\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right)} \le \sqrt{2}\sqrt{1-\lambda_{\min}^2\left(U_j^\top\widehat{U}_{j,\ell}^{(0)}\right)}$$

(4) For all $U, \widehat{U} \in \mathbb{O}^{p\times r}$,

$$\left\|U_\perp^\top\widehat{U}\right\|_2 = \left\|U_\perp^\top\widehat{U}\widehat{U}^\top\right\|_2 = \left\|U_\perp^\top UU^\top - U_\perp^\top\widehat{U}\widehat{U}^\top\right\|_2 \le \left\|UU^\top - \widehat{U}\widehat{U}^\top\right\|_2 \tag{67}$$

Since under $E(C_2)$, it holds that $\sup_{j,\ell}\left\|\widehat{U}_{j,\ell}^{(0)}\widehat{U}_{j,\ell}^{(0)\top} - U_{j,\ell}U_{j,\ell}^\top\right\|_2 \le C_2\sqrt{p}\sigma\lambda_{\min}^{-1}$, then we obtain that

$$\sup_{j,\ell}\left\|\widehat{U}_{j',\ell}^{(0)}B_{j',\ell}A_{j',\ell}^\top - U_{j'}\right\|_2 \le C_2\sqrt{2}\sqrt{p}\sigma\lambda_{\min}^{-1}$$

Combining the above inequality with (66) and using that $\|T_j\|_2 \le \kappa_0\lambda_{\min}$ and

$$\sup_{\substack{X_{j'}\in\mathbb{R}^{p_{j'}\times r_{j'}} \\ \|X_{j'}\|_2\le 1, j'\in[J]\setminus\{j\}}} \|Z_{j,\ell}(X_1\otimes X_2\otimes\cdots\otimes X_{j-1}\otimes X_{j+1}\otimes\cdots\otimes X_J)\|_2 \le C_2\sigma\sqrt{pr}$$

we obtain that

$$\sup_{j,\ell}\|\zeta_{j,\ell,2}\|_2 \le \sup_{j,\ell}\left\|T_j\left[\widehat{U}^{(0)}_{-j,\ell}\widehat{U}^{(0)\top}_{-j,\ell} - U_{-j}U_{-j}^\top\right]Z_{j,\ell}^\top\right\|_2 \lesssim \kappa_0 p\sigma^2\sqrt{r}$$

and

$$\sup_{j,\ell}\|\zeta_{j,\ell,4}\|_2 \le \sup_{j,\ell}\left\|Z_{j,\ell}\left[\widehat{U}^{(0)}_{-j,\ell}\widehat{U}^{(0)\top}_{-j,\ell} - U_{-j}U_{-j}^\top\right]Z_{j,\ell}^\top\right\|_2 \lesssim p^{3/2}r\sigma^3\lambda_{\min}^{-1}$$

Moreover, since

$$\begin{aligned}
&\widehat{U}^{(0)}_{-j,\ell}\widehat{U}^{(0)\top}_{-j,\ell} - U_{-j}U_{-j}^\top \\
&= \sum_{j'\neq j}\Bigg[\left(U_1U_1^\top\right)\otimes\cdots\otimes\left(U_{j'-1}U_{j'-1}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell} - U_{j'}U_{j'}^\top\right)\otimes\left(\widehat{U}^{(0)}_{j'+1,\ell}\widehat{U}^{(0)\top}_{j'+1,\ell}\right) \\
&\qquad\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{j-1,\ell}\widehat{U}^{(0)\top}_{j-1,\ell}\right)\otimes\left(\widehat{U}^{(0)}_{j+1,\ell}\widehat{U}^{(0)\top}_{j+1,\ell}\right)\otimes\cdots\otimes\left(\widehat{U}^{(0)}_{J,\ell}\widehat{U}^{(0)\top}_{J,\ell}\right)\Bigg]
\end{aligned}$$

and $T_j = U_jG_jU_{-j}^\top$, we have that

$$\begin{aligned}
&T_j\left[\widehat{U}^{(0)}_{-j,\ell}\widehat{U}^{(0)\top}_{-j,\ell} - U_{-j}U_{-j}^\top\right]T_j^\top \\
&= U_jG_j\sum_{j'\neq j}\Bigg[(I_{r_1})\otimes\cdots\otimes\left(I_{r_{j'-1}}\right)\otimes\left(U_{j'}^\top\widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell}U_{j'} - I_{r_{j'}}\right)\otimes\left(U_{j'+1}^\top\widehat{U}^{(0)}_{j'+1,\ell}\widehat{U}^{(0)\top}_{j'+1,\ell}U_{j'+1}\right) \\
&\qquad\otimes\cdots\otimes\left(U_{j-1}^\top\widehat{U}^{(0)}_{j-1,\ell}\widehat{U}^{(0)\top}_{j-1,\ell}U_{j-1}\right)\otimes\left(U_{j+1}^\top\widehat{U}^{(0)}_{j+1,\ell}\widehat{U}^{(0)\top}_{j+1,\ell}U_{j+1}\right)\otimes\cdots\otimes\left(U_J^\top\widehat{U}^{(0)}_{J,\ell}\widehat{U}^{(0)\top}_{J,\ell}U_J\right)\Bigg]G_j^\top U_j^\top
\end{aligned}$$

Note that

$$U_{j'}^\top\widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell}U_{j'} - I_{r_{j'}} = U_{j'}^\top\left(\widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell} - I_{r_{j'}}\right)U_{j'} = U_{j'}^\top\widehat{U}^{(0)}_{j',\ell,\perp}\widehat{U}^{(0)\top}_{j',\ell,\perp}U_{j'}$$

and by (67),

$$\left\|U_{j'}^\top\widehat{U}^{(0)}_{j',\ell,\perp}\widehat{U}^{(0)\top}_{j',\ell,\perp}U_{j'}\right\|_2 \le \left\|U_{j'}^\top\widehat{U}^{(0)}_{j',\ell,\perp}\right\|_2^2 \le \left\|U_{j'}U_{j'}^\top - \widehat{U}^{(0)}_{j',\ell}\widehat{U}^{(0)\top}_{j',\ell}\right\|_2^2 \lesssim p\sigma^2\lambda_{\min}^{-2}$$

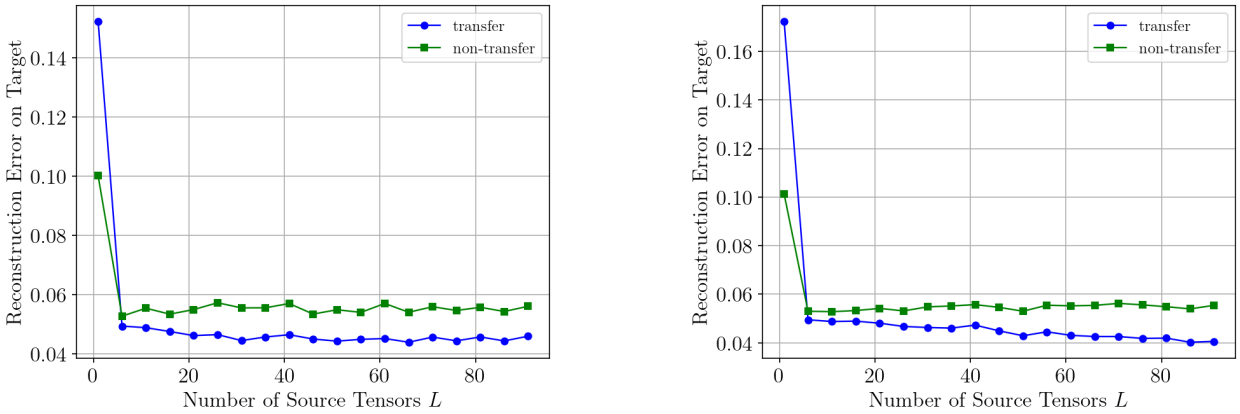
Therefore,

$$\sup_{j,\ell}\|\zeta_{j,\ell,5}\|_2 \le \sup_{j,\ell}\left\|T_j\left[\widehat{U}^{(0)}_{-j,\ell}\widehat{U}^{(0)\top}_{-j,\ell} - U_{-j}U_{-j}^\top\right]T_j^\top\right\|_2 \lesssim \kappa_0^2 p\sigma^2$$

□

D Additional Results in Numerical Studies

In the real data analysis, we further apply our transfer learning algorithm (Algorithm 3) to conduct knowledge transfer between proteins in class 0 and those in class 1. Since the sample size of class 0 is larger, we let class 0 be the source task and class 1 be the target task. Similar to the previous procedure, we randomly select L training samples on both tasks and L' test samples on the target task. The training samples on each task are averaged into a source tensor $\mathcal{T}_s$ and a target tensor $\mathcal{T}_t$ and then input into Algorithm 3 to obtain estimators $\left\{[\widehat{U}_j\ \widehat{V}_{j,t}]\right\}_{j=1,2,3}$. Specifically, we choose



(a) $\boldsymbol{r}_U = (1, 1, 1)$ (b) $\boldsymbol{r}_U = (1, 2, 2)$

Figure 6. The reconstruction errors of transfer learning between class 0 and class 1 of the PROTEINS dataset.

the ranks $\boldsymbol{r}$ to be $(2, 4, 4)$ and the ranks of common component $\boldsymbol{r}_U = (r_{1,U}, r_{2,U}, r_{3,U})$ to be $(1, 1, 1)$ or $(1, 2, 2)$. For the sample sizes, we still let $L \in [1, 100]$ and $L' = 100$. For comparison, we also record the performance of the "non-transfer" estimators, which are obtained only using the training samples on the target task. The results are displayed in Figure 6. We observe that the transfer learning method consistently outperforms the "non-transfer" method for both $\boldsymbol{r}_U = (1, 1, 1)$ and $\boldsymbol{r}_U = (1, 2, 2)$, demonstrating the advantage of leveraging knowledge from the source task.