arXiv:2007.02404v2 [stat.ME] 2 Apr 2024

Semiparametric Tensor Factor Analysis by Iteratively Projected SVD

Elynn Y. Chen[1], Dong Xia[2], Chencheng Cai[3] and Jianqing Fan[4,5]

[1]New York University, USA,

[2]Hong Kong University of Science and Technology, China

[3]Washington State University, USA,

[4] Fudan University, China; [5] Princeton University, USA

Abstract

This paper introduces a general framework of Semiparametric TEnsor Factor Analysis (STEFA) that focuses on the methodology and theory of low-rank tensor decomposition with auxiliary covariates. STEFA models extend tensor factor models by incorporating auxiliary covariates in the loading matrices. We propose an algorithm of Iteratively Projected SVD (IP-SVD) for the semiparametric estimation. It iteratively projects tensor data onto the linear space spanned by the basis functions of covariates and applies SVD on matricized tensors over each mode. We establish the convergence rates of the loading matrices and the core tensor factor. The theoretical results only require a sub-exponential noise distribution, which is weaker than the assumption of sub-Gaussian tail of noise in the literature. Compared with the Tucker decomposition, IP-SVD yields more accurate estimators with a faster convergence rate. Besides estimation, we propose several prediction methods with new covariates based on the STEFA model. On both synthetic and real tensor data, we demonstrate the efficacy of

[1]Fan is the corresponding author. Email: jqfan@princeton.edu. [1,2] Equal contribution.

the STEFA model and the IP-SVD algorithm on both the estimation and prediction tasks.

1 Introduction

Nowadays large-scale datasets in the format of matrices and tensors (or multi-dimensional arrays) routinely arise in a wide range of applications. The low-rank structure, among other specific geometric configurations, is of paramount importance to enable statistically and computationally efficient analysis of such datasets. The low-rank *tensor factor models* assume the following *noisy* Tucker decomposition:

$$\mathcal{Y} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \cdots \times_M \mathbf{A}_M + \mathcal{E}, \tag{1}$$

where $\mathcal{Y}$ is the M-th order tensor observation of dimension $I_1 \times \cdots \times I_M$, the *latent tensor factor* $\mathcal{F}$ is of dimension $R_1 \times \cdots \times R_M$, the *loading* matrix $\mathbf{A}_m$ is of dimension $I_m \times R_m$ with $R_m \ll I_m$ for each $m \in [M]$, and the noise $\mathcal{E}$ is an M-th order tensor with the same dimension of $\mathcal{Y}$. Tucker decomposition is a widely used form of tensor decomposition (Kolda and Bader, 2009; De Lathauwer et al., 2000b) and has been studied from different angles in mathematics, statistics and computer science. Particularly, the statistical and computational properties of the decomposition have been analyzed in Zhang and Xia (2018); Richard and Montanari (2014); Allen (2012a,b); Wang and Song (2017); Zhang (2019) under the general setting and in Zhang and Han (2019) under the sparsity setting where parts of the loading matrices $\{\mathbf{A}_m : m \in [M]\}$ contain row-wise sparsity structures.

Tucker decomposition to the tensor factor model is similar to singular value decomposition (SVD) to the classic vector factor model, the latter being one of the most useful tools for modeling low-rank structures in biology, psychometrics, economics, business and so on. Theoretical analyses on multivariate factor models assume i.i.d Gaussian noise at early stages (Anderson and Rubin, 1956; Anderson, 1962) and later allow for variable-wise and sample-wise correlations (Bai and Ng, 2002; Bai, 2003; Bai et al., 2012). Chapter 11 of Fan et al. (2020) and the references therein provide a thorough review of recent advances

and applications of multivariate factor models. For 2nd-order tensor (or matrix) data, Wang et al. (2019); Chen et al. (2019, 2020b) consider the matrix factor model which is a special case of (1) with $M = 2$ and propose estimation procedures based on the second moments. Later, Chen et al. (2022) extends the idea to the model (1) with arbitrary M by using the mode-wise auto-covariance matrices.

While the vanilla tensor factor model (1) is neat and fundamental, it cannot incorporate any additional information that may be relevant. Nowadays, the boom of data science has brought together informative covariates from different domains and multiple sources, in addition to the tensor observation $\mathcal{Y}$. For example, the gene expression measurements from breast tumors can be cast in a tensor format, and the relevant covariates of the cancer subtypes are usually viewed as a partial driver of the underlying patterns of genetic variation among breast cancer tumors (Schadt et al., 2005; Li et al., 2016). In restaurant recommendation system, online review sites like Yelp have access to shopping histories and friendship networks of customers, as well as the cuisine and ratings of restaurants (Acar et al., 2011). The covariate-assisted factor models have been explored for vector and matrix observations (Connor and Linton, 2007; Connor et al., 2012; Fan et al., 2016; Mao et al., 2019). Their results show that sharing relevant covariate information across datasets leads to not only a more accurate estimation but also a better interpretation.

Inspired by those prior arts, we introduce a new modeling framework – Semi-parametric TEnser Factor Analysis (STEFA) model – to leverage the auxiliary information provided by mode-wise covariates. STEFA captures practically important situations in which the observed tensor $\mathcal{Y}$ has an intrinsic low rank structure and the structure in m-th mode is *partially* explainable by some relevant covariate $\mathbf{X}_m$. The model is semi-parametric in the sense that it still allows covariate-free low-rank factors as in (1). In the special case when $\mathbf{X}_m$'s are unavailable, STEFA reduces to the classical tensor factor model (1). As to be shown in Section 6, with auxiliary covariates, our STEFA model can outperform the vanilla tensor factor model in many scenarios. The auxiliary information of $\mathbf{X}_m$ not only improves the performances of estimating latent factors but also enables prediction on new input covariates,

which is an essential difference between our proposed framework and the existing tensor decomposition literature (Richard and Montanari, 2014; Zhang and Xia, 2018; Zhang and Han, 2019; Cai et al., 2019; Sun et al., 2017; Wang and Li, 2020; Zhou et al., 2021). Indeed, unlike those tensor SVD or PCA models where estimating the latent factors usually only acts as the proxy of dimension reduction, STEFA utilizes those estimators for prediction with new observed covariates. Another popular way of incorporating auxiliary covariates information is to couple tensors and matrix covariates together for joint factorization (Acar et al., 2011; Song et al., 2019). Such method assumes that the covariate matrix and tensor share the same loading matrix along one mode. Our method is different in that auxiliary covariates can partially predict loading matrices through nonparametric function approximation. Hao et al. (2021) also used additive model in nonparametric tensor regression. But those authors dealt with tensor predictors and scalar responses, rather than a tensor of responses.

On the methodological aspect, we propose a computationally efficient algorithm, called Iteratively Projected SVD (IP-SVD), to estimate both the covariate-relevant loadings and covariate-independent loadings in STEFA. As shown in Section 4, a typical projected PCA method from Fan et al. (2016), while computationally fast, is generally sub-optimal because it ignores multi-dimensional tensor structures. The IP-SVD yields more accurate estimators of both the latent factors and loadings by adding a simple iterative projection after the initialization by projected PCA. On the other hand, the IP-SVD can be viewed as an alternating minimization algorithm which solves a constrained tensor factorization program where the low-rank factors are constrained to a certain functional space. The dimension of this functional space, based on the order of sieve approximation, can be significantly smaller than the ambient dimension which makes IP-SVD faster than the standard High-Order Orthogonal Iteration (HOOI) for solving the vanilla Tucker decomposition. As a result, IP-SVD requires also weaker signal-to-noise ratio conditions for convergence in general.

Theoretically, we discovered interesting properties of STEFA that are different from those of the vanilla tensor factor model (1). As proved in Richard and Montanari (2014); Zhang and Xia (2018), the HOOI algorithm achieves statistically optimal convergence rates for model

(1) as long as the signal-to-noise ratio SNR $\gtrsim (I_1 I_2 \cdots I_M)^{1/4}$ where the formal definition of SNR is deferred to Section 4. However, due to the constraint of a low-dimensional (compared with I_m) functional space, the SNR condition required by IP-SVD in STEFA is SNR $\gtrsim (J_1 J_2 \cdots J_M)^{1/4}$ where J_m is the number of basis function used in functional approximation and can be much smaller than I_m. Note that this weaker SNR condition is sufficient even for estimating the covariate-independent components. Surprisingly, it shows that covariate information is not only beneficial to estimating the covariate-relevant components but also to the covariate-independent components. Concerning the statistical convergence rates of IP-SVD, there are two terms which comprise of a parametric rate and a non-parametric rate. By choosing a suitable order for sieve approximation, we can obtain a typical semi-parametric convergence rate for STEFA which fills a void of understanding non-parametric ingredients of tensor factor models. On the technical front, investigating the theoretical properties of STEFA is challenging due to the iterative nature of the estimation procedure, which involves both a parametric and non-parametric component. Furthermore, our theoretical results only require a sub-exponential tail on the noise, which is weaker than the Gaussian or sub-Gaussian distributions of noise in all these prior works. This technical improvement may be of independent interests.

Notation and organization. The following notations are used throughout the paper. We use lowercase letter x, boldface letter $\mathbf{x}$, boldface capital letter $\mathbf{X}$, and calligraphic letter $\mathcal{X}$ to represent scalar, vector, matrix and tensor, respectively. We denote $[N] = \{1, \ldots, N\}$ for a positive integer N. For any matrix $\mathbf{X}$, we use $\mathbf{x}_{i\cdot}$, $\mathbf{x}_{\cdot j}$, and x_{ij} to refer to its i-th row, j-th column, and ij-th entry, respectively. All vectors are column vectors and row vectors are written as $\mathbf{x}^\top$. The set of $N \times K$ orthonormal matrices is defined as $\mathbb{O}^{N\times K}$. We denote $\sigma_i(\mathbf{X})$ as the i-th largest singular value of $\mathbf{X}$, $\|\mathbf{X}\|$ as the spectral norm of $\mathbf{X}$, i.e., $\|\mathbf{X}\| = \sigma_1(\mathbf{X})$, and $\|\mathbf{X}\|_{\mathrm{F}}$ as the Frobenius norm of $\mathbf{X}$. In addition, we frequently use the projection matrices $\mathbf{P}_X = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ and $\mathbf{P}_X^\perp = \mathbf{I} - \mathbf{P}_X$ where $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ denotes the Moore-Penrose generalized inverse.

The rest of this paper is organized as follows. Section 2 introduces the STEFA model and

a set of identification conditions. Section 3 proposes the IP-SVD algorithm to estimate the STEFA model and considers prediction with new covariates. Section 4 establishes theoretical properties of the estimators. Section 5 studies the finite sample performance via simulations. Section 6 presents empirical studies of two real data sets. All proofs and technique lemmas are relegated to the supplementary material.

2 STEFA: Semi-parametric TEnsor FActor model

In this section, we introduce the Semi-parametric TEnsor FActor (STEFA) model. We present it with third-order tensors ($M = 3$) to simply notation while the properties hold for general M. More information of tensor algebra can be found in Kolda and Bader (2009).

2.1 Tensor factor model

For a tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the mode-1 slices of $\mathcal{S}$ are matrices $\mathbf{S}_{i_1::} \in \mathbb{R}^{I_2 \times I_3}$ for any $i_1 \in [I_1]$ and the mode-1 fibers of $\mathcal{S}$ are vectors $\mathbf{s}_{:i_2 i_3} \in \mathbb{R}^{I_1}$ for any $i_2 \in [I_2]$ and $i_3 \in [I_3]$. We define its mode-1 matricization as a $I_1 \times I_2 I_3$ matrix $\mathcal{M}_1(\mathcal{S})$ such that $[\mathcal{M}_1(\mathcal{S})]_{i_1, i_2+(i_3-1)I_2} = s_{i_1 i_2 i_3}$, for all $i_1 \in [I_1], i_2 \in [I_2],$ and $i_3 \in [I_3]$. In other words, matrix $\mathcal{M}_1(\mathcal{S})$ consists of all mode-1 fibers of $\mathcal{S}$ as columns. For a tensor $\mathcal{F} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ and a matrix $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times R_1}$, the *mode*-1 *product* is a mapping defined as $\times_1 : \mathbb{R}^{R_1 \times R_2 \times R_3} \times \mathbb{R}^{I_1 \times R_1} \mapsto \mathbb{R}^{I_1 \times R_2 \times R_3}$ as $\mathcal{F} \times_1 \mathbf{A}_1 = \left[\sum_{r_1=1}^{R_1} a_{i_1 r_1} f_{r_1 r_2 r_3}\right]_{i_1 \in [I_1], r_2 \in [R_2], r_3 \in [R_3]}$. In a similar fashion, we can define fibers, mode matricization, and mode product for mode-2 and mode-3, respectively.

The widely used Tucker ranks (or multilinear ranks) of a tensor $\mathcal{S}$ is defined by the triplet $\text{rank}(\mathcal{S}) := (R_1, R_2, R_3)$ where $R_m = \text{rank}(\mathcal{M}_j(\mathcal{S}))$ for modes $m = 1, 2, 3$. The Tucker rank (R_1, R_2, R_3) is closely associated with the Tucker decomposition. If a tensor $\mathcal{S}$ has an exact tensor rank (R_1, R_2, R_3), then there exists a core tensor $\mathcal{F} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ such that $\mathcal{S}$ has a *Tucker decomposition* $\mathcal{S} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3$ where $\mathbf{A}_m \in \mathbb{R}^{I_m \times R_m}$, $m \in [3]$, are orthonormal matrices of the left singular vectors of $\mathcal{M}_m(\mathcal{S})$ respectively.

Given a tensor observation $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, a *tensor factor model* assumes that

$$\mathcal{Y} = \mathcal{S} + \mathcal{E} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3 + \mathcal{E}, \tag{2}$$

where the *latent tensor factor* $\mathcal{F}$ is of dimension $R_1 \times R_2 \times R_3$, the *loading* matrices $\mathbf{A}_m \in \mathbb{R}^{I_m \times R_m}$ are unknown deterministic parameters, and $\mathcal{E}$ is the noise tensor. The low-rank structure is captured by the assumption of $R_m \ll I_m$ along the m-th mode. Model (2) encompasses the vector and the matrix factor models as special sub-cases: the vector factor model (Fan et al., 2020) corresponds to the special case of $\mathcal{Y} = \mathbf{A}_1\mathcal{F}+\mathcal{E}$ where $\mathcal{Y}, \mathcal{E} \in \mathbb{R}^{I_1}$ and $\mathcal{F} \in \mathbb{R}^{R_1}$ are all vectors (i.e. 1st-order tensor). The matrix factor model (Wang et al., 2019; Chen et al., 2019; Chen and Fan, 2023) corresponds to the special case of $\mathcal{Y} = \mathbf{A}_1\mathcal{F}\mathbf{A}_2^\top + \mathcal{E}$ where $\mathcal{Y}, \mathcal{E} \in \mathbb{R}^{I_1 \times I_2}$ and $\mathcal{F} \in \mathbb{R}^{R_1 \times R_2}$ are all matrices (i.e. 2nd-order tensors).

All the components on the right hand side of model (2) are not directly observable, thus the tuples $\left(\mathcal{F} \times_1 \mathbf{H}_1^{-1} \times_2 \mathbf{H}_2^{-1} \times_3 \mathbf{H}_3^{-1}, \mathbf{A}_1\mathbf{H}_1, \mathbf{A}_2\mathbf{H}_2, \mathbf{A}_3\mathbf{H}_3\right)$ and $(\mathcal{F}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3)$ are indistinguishable for any invertible matrix $\mathbf{H}_m \in \mathbb{R}^{R_m \times R_m}$, $m \in [3]$. This is a common issue with latent models since they can only be identified up to the columns space of $\mathbf{A}_m$ (Bai, 2003; Zhang and Xia, 2018; Fan et al., 2020). To identify one representative matrix of the column space $\mathbf{A}_m$, we restrict our solution to the one that satisfies Assumption 1. Lemma 1 confirms the validity of Assumption 1 as an identification condition for model (2).

Assumption 1 (Tensor Factor Model Identification Condition)**.** *We restrict our estimation targets to the loading matrices and core tensor that satisfy (i)* $\mathbf{A}_m^\top\mathbf{A}_m/I_m = \mathbf{I}_{R_m}$ *for all* $m \in [M]$ *where* $\mathbf{I}_{R_m}$ *is an* $R_m \times R_m$ *identity matrix; and (ii)* $\mathcal{M}_m(\mathcal{F})\mathcal{M}_m(\mathcal{F})^\top$ *is a diagonal matrix with non-zero decreasing singular values for all* m*.*

Lemma 1. *Given an* $\mathcal{S} \in \mathbb{R}^{I_1 \times \cdots \times I_M}$ *with Tucker ranks* $(R_1, \cdots, R_M)$ *and* $\mathcal{M}_m(\mathcal{S})\mathcal{M}_m(\mathcal{S})^\top$ *having distinct non-zero singular values for all* m*, then there exist unique*[1] $\mathbf{A}_1, \cdots, \mathbf{A}_M$ *and* $\mathcal{F}$ *satisfying Assumption 1 so that* $\mathcal{S} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \cdots \times_M \mathbf{A}_M$*.*

[1]Note that uniqueness is up to column-wise signs of $\mathbf{A}_m$'s.

Model (2) can be estimated by solving the optimization program

$$\min_{\mathcal{F},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3} \|\mathcal{Y} - \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3\|_F^2, \tag{3}$$

under the constraints in Assumption 1. It is highly non-convex and computationally NP-hard. The higher order orthogonal iteration (HOOI) algorithm (De Lathauwer et al., 2000a) solves (3) by alternating minimization along the direction of $\mathbf{A}_m$. Given an initial guess of $\{\widehat{\mathbf{A}}_m\}_{m\geq 2}$, the algorithm update $\widehat{\mathbf{A}}_1$ to be the maximizing value $\widehat{\mathbf{A}}_1 = \sqrt{I_1}\cdot\mathrm{SVD}_{R_1}\big(\mathcal{M}_1(\mathcal{Y})(\widehat{\mathbf{A}}_2 \otimes \widehat{\mathbf{A}}_3)\big)$ where $\mathrm{SVD}_r(\cdot)$ returns top-r left singular vectors of a given matrix. Then, the algorithm proceeds to iteratively updating $\widehat{\mathbf{A}}_m$ while fixing the other $\widehat{\mathbf{A}}_j, j \neq m$ until some stopping criterion is satisfied. The performance of HOOI usually relies on the initial input of $\{\widehat{\mathbf{A}}_m\}_{m\in[M]}$.

One way to measure the importance of each factor dimension along a mode is through the mode-wise percentage explained variance. Suppose we are interested in the relative importance of mode-1 factors, the total variance along mode-1 can be calculated by $\sigma_1^2 = \mathrm{Tr}\left(\mathcal{M}_1(\mathcal{Y})\mathcal{M}_1(\mathcal{Y})^\top/(I_2I_3)\right)$ and variances of the R_1 factors of mode-1 are the diagonal elements in the covariance matrix $\mathbf{\Sigma}_{F,1} = \mathcal{M}_1(\mathcal{F})\mathcal{M}_1(\mathcal{F})^\top/(R_2R_3)$. Then the mode-1 percentage explained variances for each of the R_1 factors corresponds to each element in $\mathrm{diag}\left(\mathbf{\Sigma}_{F,1}\right)/\sigma_1^2$, where $\mathrm{diag}(\cdot)$ extracts R_1 diagonal elements from matrix $\mathbf{\Sigma}_{F,1}$.

2.2 Semiparametric tensor factor model

We now generalize the classic tensor factor model to integrate mode-wise auxiliary covariates. For any $i_1 \in [I_1]$, let $\mathbf{x}_{1,i_1} = [x_{1,i_1 1}, \cdots, x_{1,i_1 D_1}]^\top$ be a D_1-dimensional vector of covariates associated with the i_1-th entry along mode 1. We assume that the mode-1 loading coefficient a_{1,i_1r_1} can be (partially) explained by $\mathbf{x}_{1,i_1}$ such that

$$a_{1,i_1r_1} = g_{1,r_1}\left(\mathbf{x}_{1,i_1}\right) + \gamma_{1,i_1r_1}, \quad i_1 \in [I_1], r_1 \in [R_1],$$

where $g_{1,r_1} : \mathbb{R}^{D_1} \mapsto \mathbb{R}$ is a function and γ_{1,i_1r_1} is the part that *cannot* be explained by the covariates. Under this assumption, the entries in the i_1-th mode-1 slice, $i_1 \in [I_1]$, can be

written as

$$y_{i_1 i_2 i_3} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} \left(g_{1,r_1} \left(\mathbf{x}_{1,i_1} \right) + \gamma_{1,i_1 r_1} \right) a_{2,i_2 r_2} a_{3,i_3 r_3} f_{r_1 r_2 r_3} + \varepsilon_{i_1 i_2 i_3}, \tag{4}$$

for all $i_2 \in [I_2]$ and $i_3 \in [I_3]$. Let $\mathbf{X}_1$ be a $I_1 \times D_1$ matrix taking $\mathbf{x}_{1,i_1}^\top$ as rows, $\mathbf{G}_1(\mathbf{X}_1)$ be the $I_1 \times R_1$ matrix with its i_1-th row being $[g_{1,1}(\mathbf{x}_{1,i_1}), \cdots, g_{1,R_1}(\mathbf{x}_{1,i_1})]$, and $\mathbf{\Gamma}_1$ be the $I_1 \times R_1$ matrix of $[\gamma_{1,i_1 r_1}]$, we can write compactly $\mathbf{A}_1 = \mathbf{G}_1(\mathbf{X}_1) + \mathbf{\Gamma}_1$ and

$$\mathcal{Y} = \mathcal{F} \times_1 \left(\mathbf{G}_1 \left(\mathbf{X}_1 \right) + \mathbf{\Gamma}_1 \right) \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3 + \mathcal{E}. \tag{5}$$

This semi-parametric configuration is easily extendable to all modes of $\mathcal{Y}$. If any mode-m loading entries $a_{m,i_m r_m}$ can be partially explained by a D_m-dimensional vector $\mathbf{x}_{m,i_m}$, i.e. $a_{m,i_m r_m} = g_{m,r_m}(\mathbf{x}_{m,i_m}) + \gamma_{m,i_m r_m}$, then we have

$$\mathcal{Y} = \mathcal{F} \times_1 \left(\mathbf{G}_1 \left(\mathbf{X}_1 \right) + \mathbf{\Gamma}_1 \right) \times_2 \left(\mathbf{G}_2 \left(\mathbf{X}_2 \right) + \mathbf{\Gamma}_2 \right) \times_3 \left(\mathbf{G}_3 \left(\mathbf{X}_3 \right) + \mathbf{\Gamma}_3 \right) + \mathcal{E}, \tag{6}$$

where $\mathbf{X}_m$ is a $I_m \times D_m$ matrix taking $\mathbf{x}_{m,i_m}^\top$ as rows, $\mathbf{G}_m(\mathbf{X}_m)$ be the $I_m \times R_m$ matrix with its i_m-th row being $[g_{m,1}(\mathbf{x}_{m,i_m}), \cdots, g_{m,R_m}(\mathbf{x}_{m,i_m})]$, and $\mathbf{\Gamma}_m$ be the $I_m \times R_m$ matrix of $[\gamma_{m,i_m r_m}]$. We refer to (6) as the Semiparametric TEnsor FActor (STEFA) Model. An an illustration of model (5) is presented in Figure 1. When mode m has no covariates, we take $\mathbf{G}_m(\mathbf{X}_m) = \mathbf{0}$.

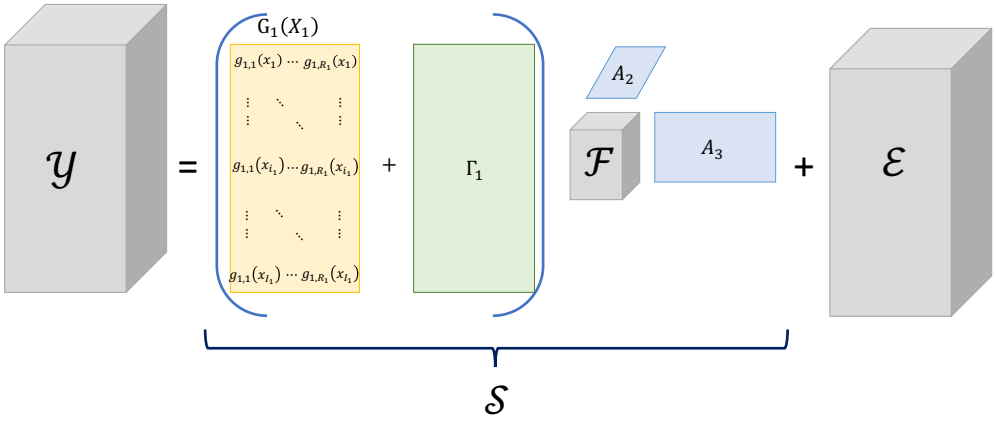


Figure 1: An illustration of the STEFA model (5).

If, additionally, mode m has no factor structure, we take $\mathbf{A}_m = \mathbf{I}_{R_m}$ – the identity matrix. If all modes have no covariates, then STEFA reduces to the classical tensor factor model (2).

STEFA is a generalization of the semi-parametric vector factor model (Fan et al., 2016) to the tensor data. But it is more complex in computation and theoretical analysis.

Remark 1 (Multivariate functional SVD)**.** *In the field of functional data analysis, researchers have studied multidimensional functional SVD (Silverman, 1996; Huang et al., 2009) and functional PCA (Zhou and Pan, 2014; Wang and Huang, 2017). Specifically, the two-way functional SVD views each entry $y_{i_1 i_2}$ of the data matrix $\mathbf{Y} \in \mathbb{R}^{I_1 \times I_2}$ as the evaluation of an underlying function $y(\cdot, \cdot)$ on a rectangular grid of sampling pints $\mathbf{x}_{1,i_1}$ and $\mathbf{x}_{2,i_2}$, that is, $y_{i_1,i_2} := y(\mathbf{x}_{1,i_1}, \mathbf{x}_{2,i_2}) := \sum_{r=1}^{R} \sigma_r g_{1,r}(\mathbf{x}_{1,i_1}) g_{1,r}(\mathbf{x}_{2,i_2})$.*

Let $\mathbf{G}_1(\mathbf{X}_1) \in \mathbb{R}^{I_1 \times R}$ be the matrix that contains $g_{1,r}(\mathbf{x}_{1,i_1})$ as its (i_1, r)-th element, $\mathbf{G}_2(\mathbf{X}_2) \in \mathbb{R}^{I_2 \times R}$ be the matrix that contains $g_{2,r}(\mathbf{x}_{2,i_2})$ as its (i_2, r)-th element, and $\mathbf{D}$ represent the diagonal matrix $\mathrm{diag}(\sigma_1, \cdots, \sigma_R)$. Under the functional SVD assumption, the data matrix has the following low-rank structure:

$$\mathbf{Y} = \mathbf{G}_1(\mathbf{X}_1)\, \mathbf{D}\, \mathbf{G}_2(\mathbf{X}_2)^\top = \mathbf{D} \times_1 \mathbf{G}_1(\mathbf{X}_1) \times \mathbf{G}_2(\mathbf{X}_2), \tag{7}$$

which is equivalent to a special case of the STEFA model where $M = 2$, core tensor $\mathcal{F} \in \mathbb{R}^{R \times R}$ is diagonal, $\mathbf{\Gamma}_1 \equiv \mathbf{0}$ and $\mathbf{\Gamma}_2 \equiv \mathbf{0}$.

The estimation method for function SVD are mostly based on regularized SVD which imposes the smoothness constraint on columns of $\mathbf{G}_1(\mathbf{X}_1)$ and $\mathbf{G}_2(\mathbf{X}_2)$. For the STEFA model, we do not impose such constraints and our projection-based algorithm also estimate the covariate independent component $\mathbf{\Gamma}_m$ that cannot be explained by the covariate.

In fact, model (7) can be extended to the higher-order setting with $M \geq 3$, which can be viewed as a functional CP tensor decomposition (Kolda and Bader, 2009) and is an interesting topic for future research.

Remark 2 (Tensor response regression)**.** *The STEFA model is related to a list of tensor response regression models (Raskutti et al., 2019) with a low-rank coefficient tensor. Notably, Sun and Li (2017) and Zhou et al. (2021) consider a model where response tensors $\mathcal{Y}_t \in \mathbb{R}^{I_1 \times \cdots \times I_{M-1}}$ are related to a D_M-dimensional vector of covariate $\mathbf{x}_t$ through*

$$\mathcal{Y}_t = \mathcal{B} \times_M \mathbf{x}_t + \mathcal{E}_t, \tag{8}$$

where $\mathcal{B}$ is a $I_1 \times \cdots \times I_{M-1} \times D_M$ unknown parameter tensor of interest, and the noise tensor $\mathcal{E}_t$ has i.i.d. standard Gaussian entries. Model (8) can be rearranged to a similar form as the STEFA model. Specifically, we stack the tensor response $\mathcal{Y}_t$ along a new M-th order and get a new $(\mathbf{I}_1 \times \cdots \times \mathbf{I}_{M-1} \times T)$ tensor $\mathcal{Y}$. We also stack the vector covariate $\mathbf{x}_t$ together and get a new $(T \times D_M)$ matrix $\mathbf{X}_M$. Then, model (8) can be rewritten as

$$\mathcal{Y} = \mathcal{B} \times_M \mathbf{X}_M + \mathcal{E}_t. \tag{9}$$

For high-dimensional data, the sparse or low-rank structure is assumed on the coefficient tensor $\mathcal{B}$ to facilitate estimation. For example, Sun and Li (2017) and Zhou et al. (2021) assume that $\mathcal{B}$ admits a rank-R CP decomposition structure. Alternatively, $\mathcal{B}$ can be assumed to admit a rank-$(R_1, \cdots, R_M)$ Tucker decomposition structure (Raskutti et al., 2019) denoted by $\mathcal{B} := \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \cdots \times_{M-1} \mathbf{A}_{M-1} \times_M \mathbf{B}_M$, where $\mathcal{F}$ is a $R_1 \times \cdots \times R_M$ tensor, $\mathbf{A}_m$ are $I_m \times R_m$ matrices for $m \in [M-1]$ and $\mathbf{B}_M$ is a $D_M \times R_M$ matrix. Under such Tucker low-rankness, model (9) can be further rewritten as

$$\mathcal{Y} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \cdots \times_{M-1} \mathbf{A}_{M-1} \times_M (\mathbf{X}_M \mathbf{B}_M) + \mathcal{E}_t.$$

which has the same form as a restricted STEFA model with $\mathbf{A}_M = \mathbf{X}_M \mathbf{B}_M$ being exact linear and non-existence of the covariate-independent component $\mathbf{\Gamma}_M$.

Remark 3 (Multiple-mode-covariate tensor regression)**.** *The multiple-mode-covariate (MMC) tensor regression (Hu et al., 2022) with identity link function writes*

$$\mathcal{Y} = \mathcal{B} \times_1 \mathbf{X}_1 \times_2 \mathbf{X}_2 \times_3 \mathbf{X}_3 + \mathcal{E}, \tag{10}$$

where $\mathbf{X}_m$ is the observable $I_m \times D_m$ covariate matrix and $\mathcal{B}$ is a low-rank regression coefficient tensor. The MMC tensor regression model is a parametric model while the the STEFA model is semi-parametric. The STEFA model is to the MMC tensor regression as the projected PCA is to the reduce-rank regression.

If we wish to make the parametric assumption that the true loading function $g_{m,r_m}(\cdot)$ is linear and no covariate-independent component, i.e. $\gamma_{m,i_m,r_m} = 0$ in (4), the STEFA model

can be rewritten in the same form as (10)*. Specifically, the loading can be explicitly written as* $\mathbf{A}_m = \mathbf{X}_m\mathbf{B}_m$ *where* $\mathbf{B}_m \in \mathbb{R}^{D_m \times R_m}$*. The STEFA model can be rewritten as*

$$\mathcal{Y} = \mathcal{F} \times_1 (\mathbf{X}_1\mathbf{B}_1) \times_2 (\mathbf{X}_2\mathbf{B}_2) \times_3 (\mathbf{X}_3\mathbf{B}_3) + \mathcal{E} = \mathcal{B} \times_1 \mathbf{X}_1 \times_2 \mathbf{X}_2 \times_3 \mathbf{X}_3 + \mathcal{E}, \tag{11}$$

where $\mathcal{B} = \mathcal{F} \times_1 \mathbf{B}_1 \times_2 \mathbf{B}_2 \times \mathbf{B}_3$*. Otherwise, the STEFA model is very different from the MCC tensor regression since it allows any smooth function* $g_{m,r_m}(\cdot)$ *and the existence of the covariate-independent component* γ_{m,i_m,r_m}*. Generally, the advantages of the STEFA model are its non-parametric modeling on the covariates as well as its weak technical assumptions.*

2.2.1 Identifiability conditions for STEFA

Similar to the tensor factor model (2), the identifiability is also an issue for STEFA. Note that the factor loading $\mathbf{A}_m$ in STEFA consists of two components $\mathbf{G}_m(\mathbf{X}_m)$ and $\mathbf{\Gamma}_m$. A naive generalization of Assumption 1 requires that

$$\mathbf{I}_{R_m} = \mathbf{A}_m^\top\mathbf{A}_m = (\mathbf{G}_m(\mathbf{X}_m) + \mathbf{\Gamma}_m)^\top(\mathbf{G}_m(\mathbf{X}_m) + \mathbf{\Gamma}_m) = \mathbf{G}_m(\mathbf{X}_m)^\top\mathbf{G}_m(\mathbf{X}_m) + \mathbf{\Gamma}_m^\top\mathbf{\Gamma}_m,$$

where we assume that $\mathbf{\Gamma}_m^\top\mathbf{G}_m(\mathbf{X}_m) = \mathbf{0}$. While the above identification is theoretically valid, such a condition imposes a constraint jointly for both the parametric and non-parametric components and introduces unnecessary difficulty into the estimating procedures. Instead, we propose the following identification condition for STEFA.

Assumption 2 (STEFA Identification Condition)**.** *We restrict our estimation targets to the loading matrices and core tensor that satisfy*

(i) $\mathbf{G}_m^\top(\mathbf{X}_m)\mathbf{G}_m(\mathbf{X}_m)/I_m = \mathbf{I}_{R_m}$ *and* $\mathbf{G}_m^\top(\mathbf{X}_m)\mathbf{\Gamma}_m = \mathbf{0}$ *for all* $m \in [M]$*.*

(ii) $\mathcal{M}_m(\mathcal{F})\mathcal{M}_m(\mathcal{F})^\top$ *is a diagonal matrix with non-zero decreasing singular values for all* $m \in [M]$*.*

Note that the identification condition $\mathbf{G}_m^\top(\mathbf{X}_m)\mathbf{G}_m(\mathbf{X}_m)/I_m = \mathbf{I}_{R_m}$ can be replaced with $\mathbf{\Gamma}_m^\top\mathbf{\Gamma}_m/I_m = \mathbf{I}_{R_m}$. We choose the first equation just for simplicity because our method starts with estimating the non-parametric component $\mathbf{G}_m(\mathbf{X}_m)$. However, if some mode

m has no covariate information, then we have to replace the identification condition with $\mathbf{\Gamma}_m^\top \mathbf{\Gamma}_m / I_m = \mathbf{I}_{R_m}$. Also note that $\mathbf{G}_m(\mathbf{X}_m)$ is the $I_m \times R_m$ matrix of $[g_{m,r_m}(\mathbf{x}_{m,i_m})]_{i_m,r_m}$, thus the identification condition is defined with respect to matrix $\mathbf{G}_m(\mathbf{X}_m)$ with a fixed I_m, not on the functional form of $g_{m,r_m}(\mathbf{x}_{m,i_m})$. Alternatively, one can consider a functional version of identification conditions on $g_{m,r_m}(\mathbf{x}_{m,i_m})$ defined on a Hilbert space consisting of all the square integrable functions. But the intricate combination of functional space and tensor structure renders the problem even more difficult and thus will not be pursued here.

3 Estimation

In this section, we present a computationally efficient Iteratively Projected SVD (IP-SVD) algorithm to estimate the STEFA model. Given the identification condition (Assumption 2), we start with estimating the non-parametric component $\mathbf{G}_m(\mathbf{X}_m)$.

3.1 Sieve approximation and basis projection

Our primary ingredient of estimating $\mathbf{G}_m(\mathbf{X}_m)$ is the sieve approximation which is a classical method in non-parametric statistics (Chen, 2007). At this moment, we assume that the latent dimensions R_1, R_2 and R_3 are known. In Section 3.3, we will discuss a method to consistently estimate R_1, R_2 and R_3 when they are unknown.

Sieve approximation relies on a set of basis functions. Take mode 1 for illustration. We denote $\{\phi_{1,j_1}(\cdot)\}_{j_1 \in [J_1]}$ as a set of basis functions on $\{f : \mathbb{R}^{D_1} \to \mathbb{R}^{I_1}\}$, which spans a complete space for $\{g_{1,r_1}(\cdot)\}_{r_1 \in [R_1]}$. Some widely-used basis functions are B-spline, Fourier series, wavelets, and polynomial series (Chen, 2007, Section 2.3). We let $\mathbf{\Phi}_1(\mathbf{X}_1)$ be the $I_1 \times J_1$ matrix whose (i_1, j_1)-th element is $\phi_{1,j_1}(\mathbf{x}_{1,i_1})$. We denote the $J_1 \times R_1$ matrix of sieve coefficients as $\mathbf{B}_1 = [\mathbf{b}_{1,1}, \cdots, \mathbf{b}_{1,R_1}]$, and the $I_1 \times R_1$ residual matrix as $\mathbf{R}_1(\mathbf{X}_1)$, consisting of approximation errors. Then, in the matrix form, we have $\mathbf{G}_1(\mathbf{X}_1) = \mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}_1)$, where $\mathbf{\Phi}_1(\mathbf{X}_1)$ can be constructed from covariates and $\mathbf{R}_1(\mathbf{X}_1)$ shall be small for a large enough J_1. To this end, the factor loading $\mathbf{A}_1$ can be written as $\mathbf{A}_1 = \mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}_1) +$

$\mathbf{\Gamma}_1$ (illustrated in the big parentheses in Figure 2.) Generalizing to other modes $m \in [M]$,

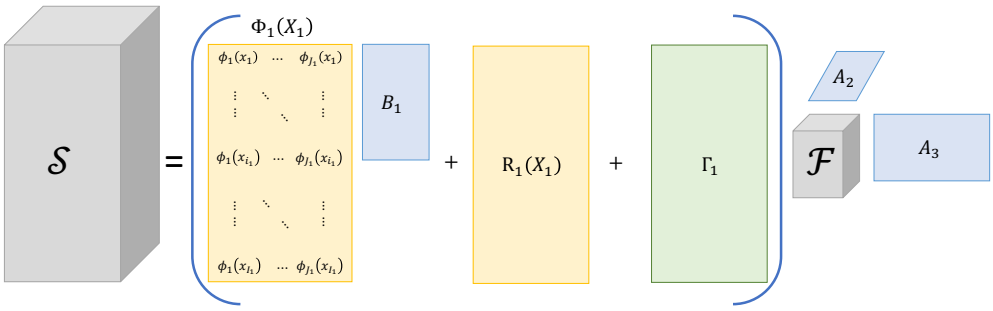


Figure 2: An illustration of sieve approximation in signal of the STEFA model. The first loading matrices $\mathbf{A}_1$ is decomposed into three part: the sieve approximation $\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1$, the sieve residual $\mathbf{R}_1(\mathbf{X}_1)$ and the covariate independent component $\mathbf{\Gamma}_1$

we can define similar terms and write that

$$\mathbf{G}_m(\mathbf{X}_m) = \mathbf{\Phi}_m(\mathbf{X}_m)\mathbf{B}_m + \mathbf{R}_m(\mathbf{X}_m). \tag{12}$$

Then, a general STEFA can be re-formulated as

$$\mathcal{Y} = \mathcal{F} \times_1 \left(\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}_1) + \mathbf{\Gamma}_1\right) \times_2 \cdots \times_M \left(\mathbf{\Phi}_M(\mathbf{X}_M)\mathbf{B}_M + \mathbf{R}_M(\mathbf{X}_M) + \mathbf{\Gamma}_M\right) + \mathcal{E}.$$

In practice, to nonparametrically estimate $g_{m,r_m}(\mathbf{x}_{m,i_m})$ without suffering from the curse of dimensionality when the dimension of $\mathbf{x}_{m,i_m}$ is large, we can assume $g_{m,r_m}(\mathbf{x}_{m,i_m})$ to be structured. A popular example of this kind is the additive model: for each $r_m \in [R_m]$, there are D_m univariate functions $\{g_{m,r_m d_m}(\cdot)\}_{d_m=1}^{D_m}$ such that

$$g_{m,r_m}(\mathbf{x}_{m,i_m}) = \sum_{d_m=1}^{D_m} g_{m,r_m d_m}(x_{m,i_m d_m}). \tag{13}$$

Each one dimensional additive component $g_{m,r_m,d_m}(x_{i_m d_m})$ can be estimated without curse of dimensionality by the sieve approximation or other more complex functions. Possible data-driven methods to estimate J's are discussed in Appendix **??** in the supplemental material.

3.2 Iteratively projected SVD

We propose an iteratively projected SVD (IP-SVD) algorithm [2] to estimate the right hand side of the STEFA model (6) from tensor $\mathcal{Y}$ and matrices of covariate $\mathbf{X}_m$ for $m \in [M]$. For ease of notations, we write $\mathbf{G}_m$ and $\mathbf{\Phi}_m$ instead of $\mathbf{G}_m(\mathbf{X}_m)$ and $\mathbf{\Phi}_m(\mathbf{X}_m)$ and define $\mathbf{P}_m = \mathbf{\Phi}_m \cdot \left(\mathbf{\Phi}_m^\top \mathbf{\Phi}_m\right)^{-1} \mathbf{\Phi}_m^\top$ as the $I_m \times I_m$ projection matrix onto the sieve spaces spanned by the basis functions of $\mathbf{X}_m$ of all $m \in [M]$. Algorithm 1 summarizes the whole procedure. For ease of presentation, it is presented for the third order tensor or $M = 3$. But it is representative for the general M setting. The outputs are estimators of the tensor factor $\widehat{\mathcal{F}}$, covariate-relevant loadings $\widehat{\mathbf{G}}_m$, sieve coefficient matrices $\widehat{\mathbf{B}}_m$, full loading matrices $\widehat{\mathbf{A}}_m$ and covariate-independent loadings $\widehat{\mathbf{\Gamma}}_m$ for all $m \in [M]$.

The algorithm is divided into two major blocks. The first block consists of the first four steps, namely projected spectral initialization, projected power iteration, projection estimate for the tensor factor and orthogonal calibration. Together, they estimate $\widehat{\mathcal{F}}$ and $\widehat{\mathbf{G}}_m$ through an iterative procedure. The first step of projected spectral initialization utilizes the fact that the column space of each loading $\mathbf{G}_m$ is mainly a subspace of the basis projection $\mathbf{P}_m$ by sieve approximation. It obtains a preliminary estimator for $\mathbf{G}_m$ for each $m \in [3]$ via sieve projection, matricization and singular value decomposition (SVD), specified in equation (14). This step, in spirit, is similar to the projected PCA in (Fan et al., 2016). This initial estimator $\widetilde{\mathbf{G}}_m^{(0)}$ acts as a good starting point, but is *sub-optimal* in general. In the second step of projected power iteration, we apply power iterations to refine the initialization. Given rudimentary estimators $\widetilde{\mathbf{G}}_2^{(t-1)}$ and $\widetilde{\mathbf{G}}_3^{(t-1)}$, we further denoise $\mathcal{Y}$ by the mode-2 and 3 projections: $\mathcal{Y} \times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top} \times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}$. This refinement can significantly reduce the amplitude of noise while reserving the mode-1 singular subspace. Iteratively for $t = 1, \cdots, t_{\max}$, we obtain an updated estimator $\widetilde{\mathbf{G}}_m^{(t)}$ for each $m \in [3]$ according to (15). This projected power iteration algorithm is a modification of the classical HOOI algorithm (De Lathauwer et al., 2000a). The additional projection $\mathbf{P}_m$ restricts the solution to be a linear function of sieve basis functions. Empirically, the projected version of HOOI in this step converges very

[2] A Python library of IP-SVD is available at https://github.com/ElynnCC/STEFA-Code.git.

fast within a few iterations. The output of this step is the final estimators $\widetilde{\mathbf{G}}_m = \widetilde{\mathbf{G}}_m^{(t_{\max})}$ for $m \in [M]$. In the third step, $\widehat{\mathcal{F}}$ is estimated via least squares, which amounts to the projection in equation (16). The fourth step fixes a numerical solution of tensor factor and loadings that satisfy Assumption 2 by orthogonal calibration. The orthogonal rotation matrices are calculated by (17) and the ultimate estimator is given by equation (18).

The second main block of the algorithm takes care of the estimation of the sieve coefficient matrices $\widehat{\mathbf{B}}_m$, full loading matrices $\widehat{\mathbf{A}}_m$, and the covariate-independent loading matrices $\widehat{\mathbf{\Gamma}}_m$ for $m \in [M]$ in the fifth and sixth steps, respectively. The sieve coefficients $\mathbf{B}_m$ is useful for prediction on new covariates. After obtaining $\widehat{\mathbf{G}}_m(\mathbf{X}_m)$, sieve coefficients can be estimated following the standard sieve approximation procedure. Indeed, we estimate $\widehat{\mathbf{B}}_m$ by equation (19). Then the mode-m loading function $\mathbf{g}_m(\mathbf{x}) = (g_{m,1}(\mathbf{x}), \cdots, g_{m,R_m}(\mathbf{x}))$ can be estimated by $\widehat{\mathbf{g}}(\mathbf{x}) = \mathbf{\Phi}(\mathbf{x})\widehat{\mathbf{B}}_m$ for any $\mathbf{x}$ in the domain of mode-m covariates. Further, with the estimated $\widehat{\mathbf{G}}_m$ and tensor factor $\widehat{\mathcal{F}}$, we estimate $\widehat{\mathbf{A}}_m$ by regression in (20) and $\widehat{\mathbf{\Gamma}}_m$ by projecting $\widehat{\mathbf{A}}_m$ on the orthogonal column space of $\mathbf{\Phi}_m$ in (21).

The above procedure only involves matrix product and matrix SVD, which computes fast. Without loss of generality, assume $I_1 \geq \cdots \geq I_M$ and $R_1 \geq \cdots \geq R_M$. The major computation load comes from the first three three steps. Specifically, the projected spectral initialization in the first step requires $O(I_1^2 I_2 \cdots I_M)$ flops; each iteration in the second step requires $O(I_1 \cdots I_M R_1 \cdots R_{M-1})$ flops; and the third step requires $O(I_1 \cdots I_M R_1 \cdots R_M)$ flops. Distributed or parallel computing can be employed to speed up the computation (De Almeida and Kibangou, 2014; Baskaran et al., 2017).

3.3 Estimating the Tucker ranks

In this section, we discuss the problem of estimating the Tucker ranks (R_1, R_2, R_3) when they are unknown. Given $\mathcal{Y} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3 + \mathcal{E}$ with the identifiable condition in Assumption 1, the mode-1 matricization of $\mathcal{Y}$ is

$$\mathcal{M}_1(\mathcal{Y}) = \mathbf{A}_1 \mathcal{M}_1(\mathcal{F})(\mathbf{A}_2 \otimes \mathbf{A}_3)^\top + \mathcal{M}_1(\mathcal{E}). \tag{22}$$

Algorithm 1: Iteratively Projected SVD (IP-SVD)

Input : Tensor $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, matrices of covariate $\mathbf{X}_m$ whose rows are $\mathbf{x}_{m,i_m}$, ranks R_m, and sets of basis functions $\{\phi_{m,j_m}(\cdot)\}_{j_m \in [J_m]}$ for $m \in [3]$.

Output: $\widehat{\mathcal{F}}$, $\widehat{\mathbf{G}}_m$, $\widehat{\mathbf{B}}_m$, $\widehat{\mathbf{A}}_m$ and $\widehat{\mathbf{\Gamma}}_m$ for $m \in [3]$.

For each $m \in [3]$, calculate the projection matrices $\mathbf{P}_m = \mathbf{\Phi}_m \cdot \left(\mathbf{\Phi}_m^\top \mathbf{\Phi}_m\right)^{-1} \mathbf{\Phi}_m^\top$, where $\mathbf{\Phi}_m(\mathbf{X}_m)$ is the $I_m \times J_m$ matrix whose (i_m, j_m)-th element is $\phi_{m,j_m}(\mathbf{x}_{m,i_m})$.

/* 1st step: Projected spectral initialization. */

Let $t = 0$ and calculate

$$\widetilde{\mathcal{Y}} = \mathcal{Y} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \times_3 \mathbf{P}_3 \quad \text{and} \quad \widetilde{\mathbf{G}}_m^{(0)} = \sqrt{I_m} \cdot \mathrm{SVD}_{R_m}(\mathcal{M}_m(\widetilde{\mathcal{Y}})). \tag{14}$$

/* 2nd step: Projected power iterations. */

for $t = 1, \ldots, t_{max}$ **do**

Calculate

$$\begin{aligned}
\widetilde{\mathbf{G}}_1^{(t)} &= \sqrt{I_1} \cdot \mathrm{SVD}_{R_1}\left(\mathbf{P}_1 \cdot \mathcal{M}_1\left(\mathcal{Y} \times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top} \times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\right)\right), \\
\widetilde{\mathbf{G}}_2^{(t)} &= \sqrt{I_2} \cdot \mathrm{SVD}_{R_2}\left(\mathbf{P}_2 \cdot \mathcal{M}_2\left(\mathcal{Y} \times_1 \widetilde{\mathbf{G}}_1^{(t)\top} \times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\right)\right), \\
\widetilde{\mathbf{G}}_3^{(t)} &= \sqrt{I_3} \cdot \mathrm{SVD}_{R_3}\left(\mathbf{P}_3 \cdot \mathcal{M}_3\left(\mathcal{Y} \times_1 \widetilde{\mathbf{G}}_1^{(t)\top} \times_2 \widetilde{\mathbf{G}}_2^{(t)\top}\right)\right).
\end{aligned} \tag{15}$$

/* 3rd step: Projection estimate for tensor factor. */

Calculate , with $\widetilde{\mathbf{G}}_j = \widetilde{\mathbf{G}}_j^{t_{max}} (j = 1, 2, 3)$,

$$\widetilde{\mathcal{F}} = (I_1 I_2 I_3)^{-1} \cdot \mathcal{Y} \times_1 \widetilde{\mathbf{G}}_1^\top \times_2 \widetilde{\mathbf{G}}_2^\top \times_3 \widetilde{\mathbf{G}}_3^\top. \tag{16}$$

/* 4th step: Orthogonal calibration. */

Calculate

$$\widehat{\mathbf{O}}_m = \mathrm{SVD}_{R_m}\left(\mathcal{M}_m(\widetilde{\mathcal{F}})\mathcal{M}_m(\widetilde{\mathcal{F}})^\top\right), \text{for each } m \in [3]. \tag{17}$$

Calculate the ultimate estimator by

$$\widehat{\mathcal{F}} = \widetilde{\mathcal{F}} \times_1 \widehat{\mathbf{O}}_1^\top \times_2 \widehat{\mathbf{O}}_2^\top \times_3 \widehat{\mathbf{O}}_3^\top \quad \text{and} \quad \widehat{\mathbf{G}}_m = \widetilde{\mathbf{G}}_m \widehat{\mathbf{O}}_m, \text{for each } m \in [3]. \tag{18}$$

/* 5th step: Covariate sieve coefficient matrices. */

Calculate

$$\widehat{\mathbf{B}}_m = \left[\mathbf{\Phi}_m^\top \mathbf{\Phi}_m\right]^{-1} \mathbf{\Phi}_m^\top \widehat{\mathbf{G}}_m. \tag{19}$$

/* 6th step: Full and covariate-independent loading matrices. */

Calculate

$$\widehat{\mathbf{Q}}_m = \mathcal{M}_m\left(\widehat{\mathcal{F}} \times_{j \neq m} (\widehat{\mathbf{G}}_j / \sqrt{I_j})\right), \quad \widetilde{\mathcal{Y}}_m = \mathcal{Y} \times_{j \neq m} \mathbf{P}_j.$$

Calculate the full loading matrices by

$$\widehat{\mathbf{A}}_m = \mathcal{M}_m\left(\widetilde{\mathcal{Y}}_m\right) \widehat{\mathbf{Q}}_m^\top \left(\widehat{\mathbf{Q}}_m \widehat{\mathbf{Q}}_m^\top\right)^{-1} / \sqrt{I_m^-} \tag{20}$$

Calculate the covariate-independent loading matrices by

$$\widehat{\mathbf{\Gamma}}_m = (\mathbf{I} - \mathbf{P}_m)\widehat{\mathbf{A}}_m \tag{21}$$

The first term in (22) is of rank R_1 when $\mathbf{A}_1 \in \mathbb{R}^{I_1 \times R_1}$ and $R_1 \ll I_1$. The second term in (22) is a $I_1 \times I_2 I_3$ noise matrix with i.i.d entries. Viewing $\mathcal{M}_1(\mathcal{F})(\mathbf{A}_2 \otimes \mathbf{A}_3)^\top$ as a whole, equation (22) is a factor model and R_1 is the corresponding unknown number of factors to be determined. There exist many approaches in consistently estimating the number of factors from the model (22). In particular, Lam and Yao (2012); Ahn and Horenstein (2013); Fan et al. (2016) proposed to estimate number of factors by selecting the largest eigenvalue ratio of $\mathcal{M}_1(\mathcal{Y})[\mathcal{M}_1(\mathcal{Y})]^\top$. Due to the noise term in (22), Fan et al. (2016) pointed out it is better to work on the projected version of $\mathcal{M}_1(\mathcal{Y})$.

Suppose $\widetilde{\mathcal{Y}} = \mathcal{Y} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \times_3 \mathbf{P}_3$ is the projected version of $\mathcal{Y}$. Then with Assumption 2, $\mathbb{E}[\mathcal{M}_1(\widetilde{\mathcal{Y}})[\mathcal{M}_1(\widetilde{\mathcal{Y}})]^\top] = I_2 I_3 \mathbf{G}_1 \mathcal{M}_1(\mathcal{F})[\mathcal{M}_1(\mathcal{F})]^\top \mathbf{G}_1^\top + \mathbb{E}[\mathbf{P}_1 \mathcal{M}_1(\mathcal{E})(\mathbf{P}_2 \otimes \mathbf{P}_3)[\mathcal{M}_1(\mathcal{E})]^\top \mathbf{P}_1^\top] = I_2 I_3 \mathbf{G}_1 \mathcal{M}_1(\mathcal{F})[\mathcal{M}_1(\mathcal{F})]^\top \mathbf{G}_1^\top + \sigma_\epsilon^2 R_2 R_3 \mathbf{P}_1 \mathbf{1}_{I_3 \times I_3} \mathbf{P}_1^\top$ has the same spectrum structure as $\mathbb{E}[\mathcal{M}_1(\mathcal{Y})[\mathcal{M}_1(\mathcal{Y})]^\top]$ but with a reduced noise term. Here σ_ϵ^2 denotes the variance of the entries of $\mathcal{E}$ and $\mathbf{1}_{I_3 \times I_3}$ is the $I_3 \times I_3$ matrix with all entries equal to one. Denote by $\lambda_k(\mathcal{M}_m(\mathcal{Y})[\mathcal{M}_m(\mathcal{Y})]^\top)$ the k-th largest eigenvalue of the mode-m matricization of the projected tensor. The eigenvalue ratio estimator of R_m is defined as

$$\widehat{R}_m = \underset{1 \leqslant k \leqslant k_{max}}{\arg\max} \frac{\lambda_k(\mathcal{M}_m(\widetilde{\mathcal{Y}})[\mathcal{M}_m(\widetilde{\mathcal{Y}})]^\top)}{\lambda_{k+1}(\mathcal{M}_m(\widetilde{\mathcal{Y}})[\mathcal{M}_m(\widetilde{\mathcal{Y}})]^\top)} \tag{23}$$

where k_{max} is an upper bound on the number of factors, such as the nearest integer of $\min\left\{I_m, \prod_{n \neq m} I_n\right\}/2$, say.

The theoretical foundation for this estimator is *partially* provided in Fan et al. (2016). Specifically, *for each mode* m, as long as there exists an $\alpha \in (0, 1]$ such that all the R_m eigenvalues of $\left(\prod_{n \neq m} I_n^{1-\alpha}\right) \mathcal{M}_m(\mathcal{F})[\mathcal{M}_m(\mathcal{F})]^\top$ are bounded between two positive constants c_{min} and c_{max}. The consistency of $\widehat{R}_m$ is provided, in terms of $\mathbb{P}[\widehat{R}_m = R_m] \to 1$, under suitable conditions (e.g., sub-Gaussian noise and $J_m = o(I_m^{1/2})$). However, while $\widehat{R}_m$ works reasonably well in simulation studies, it may be *statistically sub-optimal* for STEFA because the multi-way tensor structure is under-exploited, i.e., the low-dimensional tensor-product structure of row space of $\mathcal{M}_m(\widetilde{\mathcal{Y}})$ is ignored. A statistically more efficient approach is to also estimate $\widehat{R}_1, \cdots, \widehat{R}_M$ iteratively. The idea is similar to the iterative procedure to estimate

the loadings in the tensor factor models instead of the single-step estimation as in the vector factor models (Fan et al., 2016).

Specifically, we note that $\widetilde{\mathcal{Y}}$ in (23) is a one-time projection onto the column space of $\mathbf{P}_1, \cdots, \mathbf{P}_M$. To make the estimation of $\widehat{R}_1, \cdots, \widehat{R}_M$ iterative, $\widetilde{\mathcal{Y}}$ in (23) should be replaced by a projection of the tensor onto the column spaces of $\widetilde{\mathbf{G}}_1^{(t)}, \cdots, \widetilde{\mathbf{G}}_{m-1}^{(t)}, \widetilde{\mathbf{G}}_{m+1}^{(t-1)}, \cdots, \widetilde{\mathbf{G}}_M^{(t-1)}$ when estimating $\widehat{R}_m$ for $m = 1, \cdots, M$. However, in this case, establishing the consistency theory jointly for all $\widehat{R}_m$, i.e., $\mathbb{P}[\cap_{m=1}^M \{\widehat{R}_m = R_m\}]$ can be more challenging than that in the PCA setting (Fan et al., 2016), due to the interplay between R_m's and dependence among $\widehat{R}_m$'s. We leave the theoretical investigations of $\widehat{R}_m$ for future work and suggest interested readers to refer to a very recent work (Han et al., 2022) on the rank determination for tensor factor model.

3.4 Prediction

The STEFA model can be applied to predict unobserved outcomes from the available data. We illustrate the procedure of prediction along the first mode under model (5). Prediction along other modes can be done in a similar fashion. The task here is to predict a new $I_1^{new} \times I_2 \times I_3$ tensor $\mathcal{Y}^{new}$ with new covariate matrix $\mathbf{X}_1^{new}$ whose rows are I_1^{new} new covariate $\{\mathbf{x}_{1,i_1}^{new}\}_{i_1 \in [I_1^{new}]}$ along mode 1. Under the STEFA model (5), the tensor observation $\mathcal{Y}$ assumes the following structure

$$\mathcal{Y} = \underbrace{\mathcal{F} \times_1 \mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3}_{\text{sieve signal}} + \underbrace{\mathcal{F} \times_1 \mathbf{\Lambda}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3}_{\text{residual signal}} + \mathcal{E},$$

where $\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1$ is the part explained by the sieve approximation of $\mathbf{X}_1$ and $\mathbf{\Lambda}_1 = \mathbf{R}_1(\mathbf{X}_1) + \mathbf{\Gamma}_1$ contains the sieve residual and the orthogonal part. In Section 3, we obtain estimators $\widehat{\cdot}$ for the unknowns on the right hand side. Note that $\mathbf{\Lambda}_1$ can be estimated as a whole whereas its component $\mathbf{R}_1(\mathbf{X}_1)$ and $\mathbf{\Gamma}_1$ are not separable. With new observation $\mathbf{X}_1^{new}$, we estimate the sieve signal using

$$\widehat{\mathcal{S}}_{sieve}^{new} = \widehat{\mathcal{F}} \times_1 \mathbf{\Phi}_1(\mathbf{X}^{new})\, \widehat{\mathbf{B}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3.$$

For the residual part, we use the simple kernel smoothing over mode-1 using $\mathbf{X}_1$ and $\mathbf{X}_1^{new}$. Specifically, we have the residual signal estimator $\widehat{\mathcal{S}}_{resid} = \widehat{\mathcal{F}} \times_1 \widehat{\mathbf{\Lambda}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3$. Define the kernel weight matrix $\mathbf{W} \in \mathbb{R}^{I_1^{new} \times I_1}$ with entry

$$w_{ij} = \frac{K_h(dist(\mathbf{x}_{1,i}^{new}, \mathbf{x}_{1,j}))}{\sum_{j=1}^{I_1} K_h(dist(\mathbf{x}_{1,i}^{new}, \mathbf{x}_{1,j}))}, \quad i \in [I_1^{new}] \text{ and } j \in [I_1].$$

where $K_h(\cdot)$ is the kernel function, $dist(\cdot, \cdot)$ is a pre-defined distance function such as the Euclidean distance, and $\mathbf{x}_{1,i}$ is the i-th row of $\mathbf{X}_1$. We estimate the new residual signal by

$$\widehat{\mathcal{S}}_{resid}^{new} = \widehat{\mathcal{S}}_{resid} \times_1 \mathbf{W}, \tag{24}$$

the derivation of which is given in Section B of the supplementary material. Finally, our prediction for new entries corresponding to new covariate matrix $\mathbf{X}_1^{new}$ is given by

$$\widehat{\mathcal{Y}}^{new} = \widehat{\mathcal{S}}_{sieve}^{new} + \widehat{\mathcal{S}}_{resid}^{new}. \tag{25}$$

Remark 4. *The identification condition Assumption 2 is not restrictive in the sense that it is only used to help us separate the loadings and the factor, that is, fix a numerical solution corresponding to a specific linear transformation among multiple equivalent ones. The signal part $\mathcal{S}$ will not be affected by the specific linear transformation and thus the identification Assumption 2 will not affect the prediction. Suppose the true decomposition of the signal part $\mathring{\mathcal{S}} = \mathring{\mathcal{S}}_{sieve} + \mathring{\mathcal{S}}_{resid}$ is*

$$\mathcal{S}_{sieve} = \mathring{\mathcal{F}} \times_1 \mathbf{\Phi}_1(\mathbf{X}_1)\mathring{\mathbf{B}}_1 \times_2 \mathring{\mathbf{A}}_2 \times_3 \mathring{\mathbf{A}}_3, \quad \text{and} \quad \mathcal{S}_{resid} = \mathring{\mathcal{F}} \times_1 \mathring{\mathbf{\Lambda}}_1 \times_2 \mathring{\mathbf{A}}_2 \times_3 \mathring{\mathbf{A}}_3,$$

where $\mathring{\mathcal{F}}$, $\mathring{\mathbf{B}}_1$, $\mathring{\mathbf{A}}_2$, and $\mathring{\mathbf{A}}_3$ are the true components. Our estimation targets are restricted by Assumption 1 and 2 on observed discrete rows of $\mathbf{X}_1$ and they are linear transformations of their true counterparts. That is, $\mathbf{B}_1 := \mathring{\mathbf{B}}_1\mathbf{H}_1$, $\mathbf{A}_2 := \mathring{\mathbf{A}}_2\mathbf{H}_2$, $\mathbf{A}_3 = \mathring{\mathbf{A}}_3\mathbf{H}_3$, and $\mathcal{F} = \mathring{\mathcal{F}} \times_1 \mathbf{H}_1^{-1} \times_2 \mathbf{H}_2^{-1} \times_3 \mathbf{H}_3^{-1}$ for some invertible matrices $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$. Algorithm 1 outputs one specific solution $\widehat{\mathcal{F}}, \widehat{\mathbf{B}}_1, \widehat{\mathbf{\Gamma}}_1, \widehat{\mathbf{A}}_2, \widehat{\mathbf{A}}_3$ such that Assumption 2 is satisfied on the observed $\mathbf{X}_1$ for $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{\Gamma}}_1$, and Assumption 1 is satisfied for $\widehat{\mathbf{A}}_2$ and $\widehat{\mathbf{A}}_3$.

In Section 4, our theoretical results show that the estimators output by the Algorithm 1

is close to the estimation targets that satisfy Assumption 1 and 2. As a result, $\widehat{\mathbf{B}}_1 \approx \mathring{\mathbf{B}}_1\mathbf{H}_1$, $\widehat{\mathbf{A}}_2 \approx \mathring{\mathbf{A}}_2\mathbf{H}_2$, $\widehat{\mathbf{A}}_3 \approx \mathring{\mathbf{A}}_3\mathbf{H}_3$, *and* $\widehat{\mathcal{F}} \approx \mathring{\mathcal{F}} \times_1 \mathbf{H}_1^{-1} \times_2 \mathbf{H}_2^{-1} \times_3 \mathbf{H}_3^{-1}$. *For a new observation* $\mathbf{X}_1^{new}$, *we have*

$$\begin{aligned}
\widehat{\mathcal{S}}_{sieve}^{new} &= \widehat{\mathcal{F}} \times_1 \mathbf{\Phi}_1(\mathbf{X}_1^{new})\widehat{\mathbf{B}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3 \\
&\approx (\mathring{\mathcal{F}} \times_1 \mathbf{H}_1^{-1} \times_2 \mathbf{H}_2^{-1} \times_3 \mathbf{H}_3^{-1}) \times_1 (\mathbf{\Phi}_1(\mathbf{X}_1^{new})\mathring{\mathbf{B}}_1\mathbf{H}_1) \times_2 (\mathring{\mathbf{A}}_2\mathbf{H}_2) \times_3 (\mathring{\mathbf{A}}_3\mathbf{H}_3) \\
&= \mathring{\mathcal{F}} \times_1 \mathbf{\Phi}_1(\mathbf{X}_1^{new})\mathring{\mathbf{B}}_1\mathbf{H}_1\mathbf{H}_1^{-1} \times_2 (\mathring{\mathbf{A}}_2\mathbf{H}_2\mathbf{H}_2^{-1}) \times_3 (\mathring{\mathbf{A}}_3\mathbf{H}_3\mathbf{H}_3^{-1}) \\
&= \mathring{\mathcal{F}} \times_1 \mathbf{\Phi}_1(\mathbf{X}_1^{new})\mathring{\mathbf{B}}_1 \times_2 \mathring{\mathbf{A}}_2 \times_3 \mathring{\mathbf{A}}_3 \\
&= \mathring{\mathcal{S}}_{sieve}^{new}.
\end{aligned}$$

Here, the linear transformations $\mathbf{H}_1$, $\mathbf{H}_2$ *and* $\mathbf{H}_3$ *will depend on* $\mathbf{X}_1$. *But the key point here is that the respective* $\mathbf{H}_1$ *and* $\mathbf{H}_1^{-1}$ *transformation of* $\mathring{\mathcal{F}}$ *and* $\mathring{\mathbf{B}}_1$ *will canceled out and the signal part as a whole will not be affected by any specific* $\mathbf{H}_1$ *or* $\mathbf{X}_1$.

Remark 5 (Comparison to the MMC tensor regression)**.** *IP-SVD aims to estimate both the covariate-explainable and covariate-orthogonal components in the STEFA model while the objective of the MMC tensor regression (Xu et al., 2019; Hu et al., 2022) is to estimate the reduced-rank coefficients in a tensor regression with observed independent variables. For prediction, the covariate-explainable component in the STEFA model is predicted by Sieve approximation and the covariate-orthogonal component is predicted by kernel approximation, which are very different from the regression-based prediction in Xu et al. (2019). We report a simulation in Appendix section* **??** *to show cases when STEFA performance better in prediction.*

4 Theoretical Results

In this section, we establish the statistical properties of the estimators in Algorithm 1 assuming data is generated from model (6). Lemma 2 and Corollary 1 provide error bounds of the column spaces spanned by $\widetilde{\mathbf{G}}_m^{(t)}$ for $0 \leq t \leq t_{max}$, which concerns with the estimation errors of the projected spectral initialization and the projected power iterations in Algorithm

1. Theorem 1 provides the estimation errors of the final estimators of $\widehat{\mathbf{G}}_m$ and $\widehat{\mathcal{F}}$ from their respective estimation targets $\mathbf{G}_m$ and $\mathcal{F}$ that satisfy identification condition Assumption 2. Theorem 2 provides the errors for the covariate-independent loadings $\widehat{\mathbf{\Gamma}}_m$. We provide discussions after each theorem, revealing some interesting observations in the interaction of parametric, non-parametric estimations, and iterative tensor projection.

We impose two assumptions, respectively, on the smoothness of the loading functions and on tail behavior of the noise. The smoothness assumption is standard in the non-parametric literature, while the tail condition is weaker than what is usually assumed in the tensor decomposition literature.

Assumption 3 (Smooth loading functions)**.** *We assume that, for all tensor modes $m \in [M]$,*

(i) The loading functions $g_{m,r_m}(\mathbf{x}_m)$, $\mathbf{x}_m \in \mathcal{X}_m \in \mathbb{R}^{D_m}$ belong to a Hölder class $\mathcal{A}_c^\tau(\mathcal{X}_m)$ (τ-smooth) defined by

$$\mathcal{A}_c^\tau(\mathcal{X}_m) = \left\{ g \in \mathcal{C}^q(\mathcal{X}_m) : \sup_{[\eta]\le q} \sup_{\mathbf{x}\in\mathcal{X}_m} |D^\eta g(\mathbf{x})| \le c, \text{ and } \sup_{[\eta]=q} \sup_{\mathbf{u},\mathbf{v}\in\mathcal{X}_m} \frac{|D^\eta g(\mathbf{u}) - D^\eta g(\mathbf{v})|}{\|\mathbf{u}-\mathbf{v}\|_2^\beta} \le c \right\},$$

for some positive number c, where $\tau = q + \beta$ is assumed $\tau \ge 2$. Here, $\mathcal{C}^q(\mathcal{X}_m)$ is the space of all q-times continuously differentiable real-value functions on $\mathcal{X}_m$. The differential operator D^η is defined as $D^\eta = \frac{\partial^{[\eta]}}{\partial x_1^{\eta_1}\cdots\partial x_{d_m}^{\eta_{D_m}}}$ and $[\eta] = \eta_1 + \cdots + \eta_{D_m}$ for non-negative integers $\eta_1, \cdots, \eta_{D_m}$.

(ii) The sieve coefficients $\mathbf{b}_{m,r_m} = [b_{m,r_m,1}\ b_{m,r_m,2}\ \cdots\ b_{m,r_m,J_m}]^\top$ for all $1 \le r_m \le R_m$, satisfy, as $J_m \to \infty$,

$$\sup_{\mathbf{x}\in\mathcal{X}_m} \left| g_{m,r_m}(\mathbf{x}) - \sum\nolimits_{j=1}^{J_m} b_{m,r_m,j}\phi_j(\mathbf{x}) \right|^2 = O(J_m^{-\tau})$$

where $\{\phi_j(\cdot)\}_{j=1}^{J_m}$ is a set of basis functions, and J_m is the sieve dimension.

Assumption 3 imposes mild conditions on loading functions so that their sieve approximation errors are well controlled. It is satisfied if the loading functions $g_{m,r_m}(\mathbf{x}_m)$, $m \in [M]$, belong to the Hölder class (Tsybakov, 2008). The basis functions that satisfy Assumption 3

include polynomial, wavelet basis, and B-splines (Chen, 2007). To nonparametrically estimate $g_{m,r_m}(\mathbf{x}_m)$ without the curse of dimensionality when $\mathbf{x}_m$ is multivariate, we could impose certain low-dimensional structure on $g_{m,r_m}(\mathbf{x}_m)$, such as an additive structure used in Fan et al. (2016). To emphasis the main theoretical founding, we use τ as a given parameter in the following theorems and avoid dissecting it from the perspective of nonparametric estimation.

Assumption 4 (Sub-exponential noise)**.** *Each entry ε_ω of the noise tensor $\mathcal{E}$ are i.i.d. sub-exponential random variables with $\mathbb{E}(\varepsilon_\omega) = 0$ and $\mathbb{E}\exp(\varepsilon_\omega/K_0) \leq e$ for some constant $K_0 = O(1)$, for all $\omega \in [I_1] \times [I_2] \times [I_3]$.*

The independence condition in Assumption 4 is standard for the statistical analysis of tensor factor model (Richard and Montanari, 2014; Zhang and Xia, 2018; Xia and Zhou, 2019; Han et al., 2020) and tensor time series (Chen et al., 2022; Han et al., 2020). However, all these prior works assume the Gaussian or sub-Gaussian distributions of noise. Our Assumption 4 is weaker, which requires only a sub-exponential tail on the noise. Note that Assumption 4 implies that $\mathrm{Var}(\varepsilon_\omega) = O(1)$.

We first present the estimation errors related to the iterates of covariate-relevant loadings $\widetilde{\mathbf{G}}_m^{(t)}$ for $0 \leq t \leq t_{max}$, which correspond to the rates of convergence of the eigen-space spanned by the columns of $\mathbf{G}_m^{(t)}$. For a clear presentation, the theorems are presented for the case of $M = 3$. The results can be easily extended to higher order tensors with $M > 3$. Recall that we write I_m's for the tensor dimensions, J_m's for the sieve dimensions of covariate-relevant component, and R_m's for the Tucker ranks of covariate-independent component. We also assume that $I_1 \geq I_2 \geq I_3$ and $R_1 \geq R_2 \geq R_3$ for brevity of notations. The *signal strength* of $\mathcal{F}$ is measured by $\lambda_{\min} := \min_{m\in[M]} \sigma_{R_m}\big(\mathcal{M}_m(\mathcal{F})\big)$, which is the smallest singular value of all the matricizations of $\mathcal{F}$. The condition number of $\mathcal{F}$ is defined as $\kappa_0 := \max_{m\in[M]} \|\mathcal{M}_m(\mathcal{F})\|/\lambda_{\min}$. Since the noise has a bounded variance under Assumption 4, the signal strength $\lambda_{\min}$ is regarded as the *signal-to-noise ratio* (SNR). See a similar definition in Zhang and Xia (2018).

Lemma 2 (Projected spectral initialization and projected power iterations)**.** *Suppose that Assumptions 3 and 4 hold under model (6), the condition number $\kappa_0 = O(1)$, $J_1 \asymp J_2 \asymp J_3$,*

and $R_m \le J_m$. *If* $\sqrt{I_1 I_2 I_3}\lambda_{\min} \ge C_1\Big(\kappa_0\sqrt{R_1 J_1}\log^2 I_1 + (R_1 J_1 J_2 J_3)^{1/4}\log^2 I_1\Big)$ *and* $R_m J_m^{-\tau} \le C_1^{-1}$ *for some large enough absolute constant* $C_1 > 0$. *Then it holds with probability at least* $1 - 7I_1^{-2}$ *that*

$$I_m^{-1}\left\|\widetilde{\mathbf{G}}_m^{(0)}\widetilde{\mathbf{G}}_m^{(0)\top} - \mathbf{G}_m\mathbf{G}_m^\top\right\|_{\mathrm{F}} \le C_4\left(\frac{\sqrt{R_1 J_1}\log^2 I_1}{\lambda_{\min}\sqrt{I_1 I_2 I_3}} + \frac{\sqrt{R_1 J_1 J_2 J_3}\log^4 I_1}{\lambda_{\min}^2 I_1 I_2 I_3} + \sqrt{R_m}J_m^{-\tau/2}\right) \tag{26}$$

for some absolute constant $C_4 > 0$. *Moreover, for all* $t = 1, \cdots, t_{\max}$, *it holds with probability at least* $1 - 48I_1^{-2}$,

$$\begin{aligned}\max_m\ I_m^{-1}\left\|\widetilde{\mathbf{G}}_m^{(t)}\widetilde{\mathbf{G}}_m^{(t)\top} - \mathbf{G}_m\mathbf{G}_m^\top\right\|_{\mathrm{F}} \le \frac{1}{2}\cdot\max_m\ I_m^{-1}\left\|\widetilde{\mathbf{G}}_m^{(t-1)}\widetilde{\mathbf{G}}_m^{(t-1)\top} - \mathbf{G}_m\mathbf{G}_m^\top\right\|_{\mathrm{F}} \\ +2\sqrt{R_1}J_1^{-\tau/2} + C_4'\frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}},\end{aligned} \tag{27}$$

where $C_4' > 0$ *is an absolute constant. Therefore, after* $t_{\max} = O(\log(\lambda_{\min}\sqrt{I_1I_2I_3/J_1}) + \tau\cdot\log(J_1) + 1)$ *iterations, it holds with probability at least* $1 - 48I_1^{-2}$.

$$\max_m\ I_m^{-1}\left\|\widetilde{\mathbf{G}}_m^{(t_{\max})}\widetilde{\mathbf{G}}_m^{(t_{\max})\top} - \mathbf{G}_m\mathbf{G}_m^\top\right\|_{\mathrm{F}} \le C_5'\frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}} + 2\sqrt{R_1}J_1^{-\tau/2}, \tag{28}$$

where $C_5' > 0$ *is an absolute constant.*

Recall that τ characterizes the smoothness of the covariate-relevant loading functions. As shown in Lemma 2, if τ is larger, the estimation error decreases. The projected initialization $\widetilde{\mathbf{G}}_m^{(0)}$ in Algorithm 1 is obtained by the projected PCA (Fan et al., 2016). By Lemma 2, a warm initialization satisfying $I_m^{-1}\left\|\widetilde{\mathbf{G}}_m^{(0)}\widetilde{\mathbf{G}}_m^{(0)\top} - \mathbf{G}_m\mathbf{G}_m^\top\right\| \le 1/2$ is guaranteed as long as $\sqrt{I_1I_2I_3}\lambda_{\min} \ge C_4'(J_1J_2J_3)^{1/4}\log^2 I_1 + C_5'\sqrt{J_1}\log^2 I_1$ for $R_1 = O(1)$ and some absolute constants C_4' and C_5'. Compared with the vanilla spectral initialization (Zhang and Xia, 2018; Xia and Zhou, 2019; Richard and Montanari, 2014) that requires $\sqrt{I_1I_2I_3}\lambda_{\min} \gg (I_1I_2I_3)^{1/4}$ and sub-Gaussian noise, our projected spectral initialization requires a substantially weaker condition on the signal strength when $J_m \ll I_m$. The logarithmic factors in Lemma 2 emerge from the sub-exponential tail of noise distribution, which has never been studied in existing literature. Moreover, the initialization error (26) has two leading terms. When the signal strength $\lambda_{\min}$ is only medium strong, that is, $\sqrt{I_1I_2I_3}\lambda_{\min} \gg (J_1J_2J_3)^{1/4}$ but

$\sqrt{I_1 I_2 I_3}\lambda_{\min} \ll J_1$, the second term in (26) dominates and the initialization error is at the order of $(R_1 J_1 J_2 J_3)^{1/2}\log^4(I_1)/(\lambda_{\min}^2 I_1 I_2 I_3)$.

The initialization obtained by projected PCA (Fan et al., 2016) is sub-optimal for tensor data and the IP-SVD refines it by projected power iteration. Equation (27) shows that the error is decreasing after each mid-step projected power iteration. In the end, the error (28) of the final estimator is at a smaller order of $(J_1 R_1)^{1/2}\log^2(I_1)/\big(\lambda_{\min}(I_1 I_2 I_3)^{1/2}\big)$.

The estimation error in (28) of the final estimator is a mixture of two terms. The first term can be viewed as a parametric rate and is related to the model complexity in approximating $\mathbf{G}_m$ by the column space of $\mathbf{\Phi}(\mathbf{X}_m)$. Similar to usual parametric settings, the dimension of the $J_1 \times R_1$ parameter matrix $\mathbf{B}_1$ appears in the numerator of this first term. An interesting fact is that the parametric estimation error decreases when the signal strength $\lambda_{\min}$ increases, and increases when the Sieve dimension J_m increases.

The second term in the estimation error of (28) can be viewed as a non-parametric rate and is related to functional approximation errors which relies crucially on the Sieve dimension. This rate is unaffected when signal strength $\lambda_{\min}$ changes, but decreases when the Sieve dimension J_m increases. So there is a trade-off in choosing Sieve dimension in order to balance the parametric and non-parametric rates. The following result establishes the estimation error with the optimally-chosen Sieve dimension J_m.

Corollary 1. *Under the conditions of Lemma 2 and* $J_1 = \lceil C_6\big(\log^2(I_1)/(\lambda_{\min}\sqrt{I_1 I_2 I_3})\big)^{-2/(\tau+1)}\rceil$, *it holds that, for some absolute constants* $C_6, C_7, C_8 > 0$, *with probability at least* $1 - 48 I_1^{-2}$,

$$\max_m \; I_m^{-1}\Big\|\widetilde{\mathbf{G}}_m^{(t_{\max})}\widetilde{\mathbf{G}}_m^{(t_{\max})\top} - \mathbf{G}_m\mathbf{G}_m^\top\Big\|_{\mathrm{F}} \le C_7\sqrt{R_1}\Big(\frac{\log^2 I_1}{\lambda_{\min}\sqrt{I_1 I_2 I_3}}\Big)^{\frac{\tau}{\tau+1}} + C_8\frac{\sqrt{R_1 R_2 R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1 I_2 I_3}}.$$

The first rate in Corollary 1 dominates whenever $R_2, R_3 = O(1)$. This rate is very typical in non-parametric regression (Chen, 2007; Tsybakov, 2008) and it shows that the estimation error of $\widetilde{\mathbf{G}}_m^{(t_{\max})}$ decreases when the true loading functions are smoother in Assumption 3.

Till now, we have shown that the space spanned by the columns of the loadings $\mathbf{G}_m$ can be consistently estimated. Next, we show that the columns of $\mathbf{G}_m$ and tensor factor $\mathcal{F}$ can be determined up to a sign for the restricted estimation targets $\mathbf{G}_m$ and $\mathcal{F}$ that satisfy the

identification condition Assumption 2. The concept of the eigengap of tensor $\mathcal{F}$ is needed before we present those results. Here, we define

$$\mathrm{Egap}(\mathcal{F}) = \min_{1\le m\le M}\Big\{\min_{1\le j\le R_m}\sigma_j\big(\mathcal{M}_m(\mathcal{F})\big) - \sigma_{j+1}\big(\mathcal{M}_m(\mathcal{F})\big)\Big\},$$

where we denote $\sigma_{R_m+1}\big(\mathcal{M}_m(\mathcal{F})\big) = 0$. Intuitively, $\mathrm{Egap}(\mathcal{F})$ represents the smallest gap of singular values of $\mathcal{M}_m(\mathcal{F})$ for all $m \in [M]$. The eigengap condition on $\mathrm{Egap}(\mathcal{F})$ is imposed to ensure that the order of singular values will not be violated by small perturbations.

Theorem 1 (Covariate-relevant loadings and tensor factor)**.** *Suppose that the signal strength satisfies* $\sqrt{I_1I_2I_3}\lambda_{\min} \ge C_0\Big(\kappa_0\sqrt{R_1J_1}\log^2 I_1 + (R_1J_1J_2J_3)^{1/4}\log^2 I_1\Big)$ *under model (6), the conditions of Lemma 2 and*

$$\mathrm{Egap}(\mathcal{F}) \ge C_1\sqrt{J_1R_1^2 + R_1^2R_2R_3}\log^2(I_1)/\sqrt{I_1I_2I_3} + C_2\lambda_{\min}R_1J_1^{-\tau/2}$$

hold for some absolute constants $C_0, C_1, C_2 > 0$*. Let* $\widehat{\mathcal{F}}$ *and* $\widehat{\mathbf{G}}_m$ *be the estimators after orthogonality calibration* (18)*. Then there exist diagonal matrices* $\{\mathbf{S}_m\}_{m\in[3]}$ *whose diagonal entries are either* -1 *or* $+1$ *such that, with probability at least* $1 - 49I_1^{-2}$*,*

$$\max_{m\in[3]} I_m^{-1/2}\Big\|\widehat{\mathbf{G}}_m - \mathbf{G}_m\mathbf{S}_m\Big\|_{\mathrm{F}} \le C_7\frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}} + C_8\sqrt{R_1}J_1^{-\tau/2}$$

and

$$\|\widehat{\mathcal{F}} - \mathcal{F}\times_1\mathbf{S}_1\times_2\mathbf{S}_2\times_3\mathbf{S}_3\|_{\mathrm{F}} \le C_7'\frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\sqrt{I_1I_2I_3}} + C_8'\lambda_{\min}\sqrt{R_1}J_1^{-\tau/2}$$

where $C_7, C_8, C_7', C_8' > 0$ *are absolute constants.*

Here the columns of factor loadings $\mathbf{G}_m$ can be determined up to a sign which is common in matrix singular value decomposition. Similarly to Corollary 1, if we choose $J_1 \asymp \lceil\big(\log^2(I_1)/(\lambda_{\min}\sqrt{I_1I_2I_3})\big)^{-2/(\tau+1)}\rceil$, Theorem 1 implies that

$$\|\widehat{\mathcal{F}} - \mathcal{F}\times_1\mathbf{S}_1\times_2\mathbf{S}_2\times_3\mathbf{S}_3\|_{\mathrm{F}} \le C_7'\sqrt{R_1}\lambda_{\min}^{\frac{1}{\tau+1}}\cdot\Big(\frac{\log^2 I_1}{\sqrt{I_1I_2I_3}}\Big)^{\frac{\tau}{\tau+1}} + C_8\Big(\frac{R_1R_2R_3\log^4 I_1}{I_1I_2I_3}\Big)^{1/2}.$$

The second term on the right hand side is negligible if $\lambda_{\min}\sqrt{I_1I_2I_3} \ge C_8'(R_2R_3)^{\frac{\tau+1}{2}}\log^2 I_1$. So the first term dominates when $R_1 = O(1)$. Moreover, the first term decreases when

τ increases implying that the core tensor can be more accurately estimated if the loading functions are smoother.

Finally, we bound the estimation error for the covariate-independent components.

Theorem 2 (Covariate-independent loadings)**.** *Suppose the conditions of Theorem 1 hold under model (6). Then, for all $m = 1, 2, 3$, it holds with probability at least $1 - 50I_1^{-2}$ that*

$$\begin{aligned}\left\|\widehat{\mathbf{\Gamma}}_m - \mathbf{\Gamma}_m \mathbf{S}_m\right\|_{\mathrm{F}} \leq \quad & C_8 \|\mathbf{\Gamma}_m\| \cdot \left(\frac{\sqrt{J_1 R_1 + R_1 R_2 R_3} \log^2 I_1}{\lambda_{\min} \sqrt{I_1 I_2 I_3}} + \sqrt{R_1} J_1^{-\tau/2} \right) \\ & + C_9 \frac{\sqrt{R_1 I_1 + J_1 R_1^2 + R_1^2 R_2 R_3} \log^{3/2} I_1}{\lambda_{\min} \sqrt{I_2 I_3}}\end{aligned}$$

where $\mathbf{S}_m$ is defined as in Theorem 1 and $C_8, C_9 > 0$ are some absolute constants. By choosing $J_1 \asymp \lceil \left(\log^2(I_1) / (\lambda_{\min} \sqrt{I_1 I_2 I_3}) \right)^{-2/(\tau+1)} \rceil$, we get, with probability at least $1 - 50 I_1^{-2}$, that

$$\begin{aligned}\left\|\widehat{\mathbf{\Gamma}}_m - \mathbf{\Gamma}_m \mathbf{S}_m\right\|_{\mathrm{F}} \leq \quad & C_8' \|\mathbf{\Gamma}_m\| \cdot \left(\sqrt{R_1} \left(\frac{\log^2 I_1}{\lambda_{\min} \sqrt{I_1 I_2 I_3}} \right)^{\frac{\tau}{\tau+1}} + \frac{\sqrt{R_1 R_2 R_3} \log^2 I_1}{\lambda_{\min} \sqrt{I_1 I_2 I_3}} \right) \\ & + C_9' \frac{\sqrt{R_1 I_1 + J_1 R_1^2 + R_1^2 R_2 R_3} \log^{3/2} I_1}{\lambda_{\min} \sqrt{I_2 I_3}}\end{aligned} \tag{29}$$

for some absolute constants $C_8', C_9' > 0$.

The error bound (29) involves two terms. The second term is similar to (except a logarithmic factor) the typical rate of tensor factor models (Zhang and Xia, 2018; Richard and Montanari, 2014) if $I_1 \geq J_1 R_1 + R_1 R_2 R_3$. However, there is a crucial difference in the STEFA model since no condition is required for $\mathbf{\Gamma}_m$ (such as orthogonality of its columns). The first term in (29) emerges from the estimation error of covariate-relevant component $\mathbf{G}_m$. For ease of exposition, assume $\|\mathbf{\Gamma}_m\|_{\mathrm{F}} \asymp R_1^{1/2} \|\mathbf{\Gamma}_m\|$ and $R_1 = O(1)$. The rate (29) yields the relative error of $\widehat{\mathbf{\Gamma}}_m$ as

$$\frac{\|\widehat{\mathbf{\Gamma}}_m - \mathbf{\Gamma}_m \mathbf{S}_m\|_{\mathrm{F}}}{\|\mathbf{\Gamma}_m\|_{\mathrm{F}}} \leq C_8' \left(\frac{\log^2 I_1}{\lambda_{\min} \sqrt{I_1 I_2 I_3}} \right)^{\frac{\tau}{\tau+1}} + C_9' \frac{\sqrt{I_1} \log^{3/2} I_1}{\lambda_{\min} \sqrt{I_2 I_3} \|\mathbf{\Gamma}_m\|}. \tag{30}$$

Therefore, the estimator $\widehat{\mathbf{\Gamma}}_m$ is consistent in relative Frobenius-norm error if $\lambda_{\min} \sqrt{I_1 I_2 I_3} \gg \log^2 I_1$ and $\lambda_{\min} \|\mathbf{\Gamma}_m\| (I_2 I_3)^{1/2} \gg I_1^{1/2} \log^{3/2} I_1$. The former condition is mild in view of the signal strength condition in Theorem 1. The latter condition relies on the magnitude of

$\|\mathbf{\Gamma}_m\|$, and it dominates the former one if $\|\mathbf{\Gamma}_m\| \leq I_1 \log^{-1/2} I_1$. Basically, if $\|\mathbf{\Gamma}_m\|$ becomes smaller, a larger signal strength $\lambda_{\min}$ is required to ensure the consistency of $\widehat{\mathbf{\Gamma}}_m$.

Comparison with HOOI. Ignoring the covariate information and assuming the orthogonality of the columns of $\mathbf{\Gamma}_m$, one can apply the higher-order orthogonal iteration (HOOI) algorithm to estimate $\mathbf{\Gamma}_m$ (additional treatments are perhaps necessary to separate $\mathbf{\Gamma}_m$ from the covariate-relevant component $\mathbf{G}_m(\mathbf{X}_m)$). It is proved in Zhang and Xia (2018) that if the SNR satisfies $\lambda_{\min}\|\mathbf{\Gamma}_1\|\|\mathbf{\Gamma}_2\|\|\mathbf{\Gamma}_3\| \geq C_0(I_1^{1/2} + (I_1I_2I_3)^{1/4})$, the HOOI algorithm outputs an estimator attaining, with high probability, a relative Frobenius-norm error rate as

$$\frac{\|\widehat{\mathbf{\Gamma}}_m^{\mathsf{HOOI}} - \mathbf{\Gamma}_m\mathbf{O}_m\|_{\mathrm{F}}}{\|\mathbf{\Gamma}_m\|_{\mathrm{F}}} \leq C_8'' \frac{\sqrt{I_1}}{\lambda_{\min}\|\mathbf{\Gamma}_1\|\|\mathbf{\Gamma}_2\|\|\mathbf{\Gamma}_3\|} \tag{31}$$

where $\mathbf{O}_m$ is an orthogonal matrix that minimizes $\|\widehat{\mathbf{\Gamma}}_m^{\mathsf{HOOI}} - \mathbf{\Gamma}_m\mathbf{O}\|_{\mathrm{F}}$. For ease of comparison, let us further assume $\|\mathbf{\Gamma}_1\| \asymp \|\mathbf{\Gamma}_2\| \asymp \|\mathbf{\Gamma}_3\|$ and $I_1 \asymp I_2 \asymp I_3$. Comparing (31) to (30), if $\|\mathbf{\Gamma}_1\| \gg I_1^{1/2}$, i.e., the covariate-independent component has a signal strength (characterized by $\|\mathbf{\Gamma}_1\|$) stronger than the covariate-relevant one (that is simply $I_1^{1/2}$ by Assumption 2) , HOOI achieves a sharper error rate than our STEFA-based estimator. On the other hand, STEFA can outperform HOOI when $\|\mathbf{\Gamma}_1\| \ll I_1^{1/2}$. Nonetheless, STEFA still enjoys a major advantage over HOOI by exploiting the covariate information. Indeed, the auxiliary covariates can potentially reduce the SNR requirement. Note that our Theorem 2 suggests that an SNR condition $\sqrt{I_1I_2I_3}\lambda_{\min} \geq C_0(J_1J_2J_3)^{1/4}$ suffices to estimate the covariate-independent component, while HOOI requires an SNR condition $\|\mathbf{\Gamma}_1\|\|\mathbf{\Gamma}_2\|\|\mathbf{\Gamma}_3\|\lambda_{\min} \geq C_0(I_1I_2I_3)^{1/4}$. Therefore, if $J_1, J_2, J_3 \ll I_1$ and $\|\mathbf{\Gamma}_1\| = O(I_1^{1/2})$, STEFA requires a weaker SNR condition.

5 Numerical studies

In this section, we use Monte Carlo simulations to assess the performances of the IP-SVD algorithm on the STEFA model under different settings. In all examples, the observation tensor $\mathcal{Y}$ is generated according to model (6), of which the dimensions of the latent tensor factor and the covariates are fixed at $R_m = R = 3$ and $D_m = D = 2$. We generate the

noise tensor $\mathcal{E}$ with each entry $\varepsilon_{i_1 i_2 i_3} \sim \mathcal{N}(0,1)$. The core tensor $\mathcal{F}$ is obtained from the core tensor of the Tucker decomposition of a $R_1 \times R_2 \times R_3$ random tensor with i.i.d. $\mathcal{N}(0,1)$ entries. The core tensor is further scaled such that $\lambda_{\min} \triangleq \min_m \sigma_{R_m}(\mathcal{M}_m(\mathcal{F})) = (I_{min})^{\alpha}$, where $I_{min} = \min\{I_1, I_2, I_3\}$ with some desired value of α. This characterization of signal strength was proposed in Zhang and Xia (2018) and we focus on the low signal-to-noise ratio regime ($\alpha \leqslant 0.5$), where HOOI is known to have unsatisfactory performance.

The explanatory variable matrix $\mathbf{X}_m \in \mathbb{R}^{I_m \times D_m}$ is generated from independent uniform distribution $\mathcal{U}(0,1)$. We generate $\mathbf{G}_m = [g_{m,r_m}(\mathbf{x}_{m,i_m\cdot})]_{i_m r_m}$ by:

$$g_{m,r_m}(\mathbf{x}_{m,i_m\cdot}) = \xi_{m,r_m,0} + \sum_{d_m=1}^{D_m} \sum_{j=1}^{J^*} \xi_{m,r_m,d_m,j} \kappa^{j-1} P_j(2x_{m,i_m d_m} - 1), \tag{32}$$

where $\xi_{m,r_m,0}$ and $\xi_{m,r_m,d_m,j} \sim \mathcal{N}(0,1)$, J^* is the true number of basis functions, $\kappa \in (0,1)$ is the decay coefficient to make sure convergence of sequences as J^* increases, and $P_j(\cdot)$ is the j-th Legendre polynomial defined on $[-1,1]$. Note that J^* denotes the true sieve order used in simulation and the J used in IP-SVD is not necessarily same as J^*. The generation of $\mathbf{\Gamma}_m$ will be specified later in each setting. Whenever a non-zero $\mathbf{\Gamma}_m$ is generated, we orthonormalize the columns of $\mathbf{A}_m = \mathbf{G}_m + \mathbf{\Gamma}_m$ such that $\mathbf{A}_m^\top \mathbf{A}_m$ is an identity matrix. Here we abuse Assumption 2 a little bit in order to control the signal-to-noise ratio through the magnitude of the core matrix $\mathcal{F}$. The orthonormalized $\mathbf{A}_m$ and the original one differ by a linear transformation of columns, which does not affect the Schatter q-sin θ distance.

In what follows, we vary (I_1, I_2, I_3), α, $\mathbf{G}_m$ and $\mathbf{\Gamma}_m$ to investigate the effects of different tensor dimensions, signal-to-noise ratios and semi-parametric assumptions on the accuracy of estimating factor, loadings and loading functions. For the error of estimating the loading $\mathbf{A}_m$, we report the average Schatten q-sin θ norm ($q = 2$):

$$\ell_2(\widehat{\mathbf{A}}_m) := \left\| \sin\Theta\left(\widehat{\mathbf{A}}_m, \mathbf{A}_m\right) \right\|_2, \quad m \in [3].$$

For the error of estimating the loading function $g_{m,r}(\mathbf{x})$, $m \in [M]$, we report

$$\ell(\widehat{g}_{m,r}) := \frac{\int \left|\widehat{g}_{m,r}(\mathbf{x}) - g^\star_{m,r}(\mathbf{x})\right|^2 d\mathbf{x}}{\int \left|g^\star_{m,r}(\mathbf{x})\right|^2 d\mathbf{x}}, \quad r \in [3].$$

For the error of the estimation of a tensor $\mathcal{Y}$, we report the relative mean squared error $\text{ReMSE}_{\mathcal{Y}} = \frac{\|\widehat{\mathcal{Y}}-\mathcal{Y}\|_F}{\|\mathcal{Y}\|_F}$. For the setting where all three modes share similar properties, we only report results for the 1-st mode for conciseness. All results are based on 100 replications.

Effect of growing dimensions and signal-to-noise ratio. In the first experiment, we examine the effect of growing dimension I and different values of α. We fix $R = 3$, $J = J^* = 4$, $\mathbf{\Gamma}_m = \mathbf{0}$, and set $I_1 = I_2 = I_3 = I$. We vary $I = \{100, 200, 300\}$ and $\alpha = \{0.1, 0.3, 0.5\}$. The mean and standard deviation of $\ell_2(\widehat{\mathbf{A}}_1)$ are presented in Table 1. Since $I_1 = I_2 = I_3$, we only report the Shatten's q-sin Θ-norm for $\widehat{\mathbf{A}}_1$ as similar result holds for $\widehat{\mathbf{A}}_2$ and $\widehat{\mathbf{A}}_3$. It is clear that the IP-SVD significantly improves upon HOOI in Shatten's q-sin Θ-norm (q=2) under all settings. While both IP-SVD and HOOI perform better when α increases and worse when dimension I increases, the IP-SVD is more favorably affected by increased α and less negatively affected by increased dimension I. The error in estimating $g_{m,r}(\mathbf{x})$ for the first mode $m = 1$ is reported in Table 2, where the phenomenon is the same as those for $\ell_2(\widehat{\mathbf{A}})$. The supplementary material (Chen et al., 2020a, Section C) also reports the same phenomenon for the unbalanced setting where I_1, I_2, and I_3 are different.

Table 1: The mean and standard deviation of the the average Schatten q-sin θ loss $\ell_2(\widehat{\mathbf{A}}_1)$ and $\text{ReMSE}_{\mathcal{Y}}$, from 100 replications, under varying dimensions and signal-to-noise ratio.

	α	0.1			0.3			0.5		
	I	100	200	300	100	200	300	100	200	300
IP-SVD	$\ell_2(\widehat{\mathbf{A}}_1)$	1.305 (0.138)	1.303 (0.126)	1.292 (0.169)	0.866 (0.233)	0.621 (0.205)	0.574 (0.200)	0.274 (0.068)	0.195 (0.051)	0.152 (0.038)
	$\text{ReMSE}_{\mathcal{Y}}$	2.471 (0.519)	2.382 (0.519)	2.281 (0.483)	0.934 (0.283)	0.675 (0.212)	0.588 (0.179)	0.280 (0.065)	0.195 (0.044)	0.154 (0.035)
HOOI	$\ell_2(\widehat{\mathbf{A}}_1)$	1.707 (0.012)	1.719 (0.007)	1.724 (0.004)	1.705 (0.012)	1.719 (0.006)	1.724 (0.004)	1.581 (0.189)	1.671 (0.122)	1.691 (0.162)
	$\text{ReMSE}_{\mathcal{Y}}$	7.829 (1.632)	10.368 (2.133)	11.999 (2.379)	3.330 (0.665)	3.652 (0.798)	3.987 (0.849)	1.548 (0.323)	1.576 (0.278)	1.556 (0.269)

Effect of the number of fitting basis. In this experiment, we examine the effect of different choices of the number of fitting basis J. Specifically, we fix $I_1 = I_2 = I_3 = I = 200$, $R = 3$ and set $\mathbf{\Gamma}_m = 0$. We vary SNR by changing $\alpha = 0.3, 0.5$. The loadings are simulated

Table 2: Under varying dimensions and signal-to-noise ratio, the mean and standard deviation of the function approximation loss $\ell(\widehat{g}_{m,r})$, for model $m = 1$ and $r \in [3]$, from 100 replications. This results for modes $m = 2, 3$ are similar.

I	R	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$		
		$\ell(\widehat{g}_{1,1})$	$\ell(\widehat{g}_{1,2})$	$\ell(\widehat{g}_{1,3})$	$\ell(\widehat{g}_{1,1})$	$\ell(\widehat{g}_{1,2})$	$\ell(\widehat{g}_{1,3})$	$\ell(\widehat{g}_{1,1})$	$\ell(\widehat{g}_{1,2})$	$\ell(\widehat{g}_{1,3})$
100	3	1.653	1.699	1.786	0.745	1.082	1.295	0.382	0.429	0.575
		(1.028)	(0.749)	(0.740)	(1.224)	(1.141)	(1.014)	(1.098)	(1.078)	(1.259)
200	3	1.479	1.715	1.792	0.524	0.898	1.016	0.134	0.127	0.270
		(0.861)	(0.653)	(0.682)	(1.119)	(1.193)	(1.119)	(0.669)	(0.558)	(0.900)
300	3	1.500	1.781	1.834	0.410	0.832	1.063	0.100	0.190	0.063
		(0.929)	(0.725)	(0.669)	(0.916)	(1.220)	(1.299)	(0.548)	(0.778)	(0.392)

according to the additive sieve structure as in (32) with fixed $J^* = 16$. However, in the estimation of $\widehat{\mathbf{A}}_m$, we use different numbers of sieve orders $J = 2, 4, 8, 16$. The mean and standard deviation of $\ell_2(\widehat{\mathbf{A}}_1)$ and $\text{ReMSE}_{\mathcal{Y}}$ are reported in Table 3.

A noteworthy observation is that increasing the sieve order J does not consistently enhance the performance. For both signal-to-noise strength in Table 3, $J = 16$ does not achieve the best performance among all choices of J, even though the data is simulated with order 16. This reflects well the bias and variance trade-off. On one hand, increasing sieve order J enhances the capability of $\mathbf{G}_m$ in capturing the parametric dependence between $\mathbf{A}_m$ and $\mathbf{X}_m$. On the other hand, a large order J increases the Frobenius norm of the projected noise $\|\mathcal{E} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \times_3 \mathbf{P}_3\|_F$, which may result in a reduced signal-to-noise ratio. Large value of α is more tolerant to this signal-to-noise decrease caused by large sieve order. As shown in Table 3, the minimum error is obtained at $J = 4$ when $\alpha = 0.3$, while $J = 8$ is the optimal one when $\alpha = 0.5$. These observations align with findings in the realm of semiparametric studies. For example, extensive spline bases often exhibit overfitting tendencies and are commonly employed alongside regularization techniques (Carroll and Ruppert, 2006).

Effect of the covariate-orthogonal loading. In this experiment, we examine the effect of the covariate-orthogonal loading part $\mathbf{\Gamma}_m$. To simulate nonzero $\mathbf{\Gamma}_m$ such that $\mathbf{A}_m$ satisfies the identification condition, we first generate a matrix $\mathbf{\Lambda}_m$ with each elements drawn from independent $\mathcal{N}(0, 1)$, project it to the orthogonal complement of $\mathbf{G}_m$ and normalize each

Table 3: The average spectral and Frobenius Schatten q-sin Θ loss for $\widehat{\mathbf{A}}_1$ and relative mean square errors for $\mathcal{Y}$ under various settings.

	$\ell_2(\widehat{\mathbf{A}}_1)$				$\text{ReMSE}_{\mathcal{Y}}$			
J	2	4	8	16	2	4	8	16
$\alpha = 0.3$	1.024 (0.177)	0.910 (0.195)	1.093 (0.256)	1.486 (0.173)	0.886 (0.113)	0.872 (0.182)	1.154 (0.349)	1.781 (0.432)
$\alpha = 0.5$	0.881 (0.153)	0.503 (0.116)	0.327 (0.057)	0.445 (0.101)	0.720 (0.080)	0.467 (0.073)	0.303 (0.053)	0.398 (0.102)

column. Specifically, the r-th column of $\mathbf{\Gamma}_m$ is obtained as

$$\boldsymbol{\gamma}_{m,\cdot r} = \mu \cdot (\mathbf{I} - \mathbf{P}_{\mathbf{G}_m})\boldsymbol{\lambda}_{m,\cdot r} / \left\|(\mathbf{I} - \mathbf{P}_{\mathbf{G}_m})\boldsymbol{\lambda}_{m,\cdot r}\right\|, \quad \text{for } r = 1, \ldots, R_m,$$

where $\mathbf{P}_{\mathbf{G}_m}$ is the projection matrix of $\mathbf{G}_m$ and $\boldsymbol{\lambda}_{m,\cdot r}$ is the r-th column of $\mathbf{\Lambda}_m$. We add a scaling factor $\mu \geqslant 0$ to controls the amplitude of the orthogonal part. Note that $\mathbf{A}_m = \mathbf{G}_m + \mathbf{\Gamma}_m$ generated in this way is not necessarily an orthogonal matrix. So a final QR decomposition is conducted on $\mathbf{A}_m$ to orthonormalize the columns of $\mathbf{A}_m$. Again, we note that we orthonormalize $\mathbf{A}_m$ just in order to control the overall signal-to-noise ratio. In the experiments, we fix $I_1 = I_2 = I_3 = I = 200$, $R = 3$ and $\alpha = 0.5$ and change the values of μ. The magnitude or the Frobenious norm of $\mathbf{\Gamma}_m$ is controlled through the coefficient μ. The errors under four different choices of μ's are reported in Table 4. Note that in the simulation, $\mathbf{A}_m = \mathbf{G}_m + \mathbf{\Gamma}_m$ is normalized such that the signal-to-noise ratio of the tensor $\mathcal{Y}$ can be controlled by the core tensor $\mathcal{F}$. A larger value of μ indicates a smaller norm of the projected tensor $\widetilde{\mathcal{Y}}$ and results in a decreased signal-to-noise ratio in the projected model. As demonstrated in Table 4, the error increases as μ increases.

Table 4: The means and standard deviations of $\widehat{\mathbf{A}}_1$ and $\text{ReMSE}_{\mathcal{Y}}$ under various settings.

μ	0	0.01	0.1	1.0
$\ell_2(\widehat{\mathbf{A}}_1)$	0.877 (0.101)	0.851 (0.125)	0.876 (0.117)	1.285 (0.132)
$\text{ReMSE}_{\mathcal{Y}}$	1.043 (0.274)	0.985 (0.271)	1.031 (0.296)	2.157 (0.543)

Effect of underlying $g_{m,r_m}(\cdot)$. In this experiment, we exam the potential impact of using the additive approximation (13) of $\mathbf{G}_m$. Under the setting $I_1 = I_2 = I_3 = 200$, $R = 3$, $D = 2$, $\alpha = 0.3$, $\mathbf{\Gamma}_m = 0$ and $J = J^* = 3$, we simulate $\mathbf{G}_m$ according to the additive case (32) and plot the true function $g^\star_{1,1}(\mathbf{x})$ and the estimated function $\widehat{g}_{1,1}(\mathbf{x})$ in Figure 3. As the additive assumption is valid for this case, the estimated function is pretty close the true one.

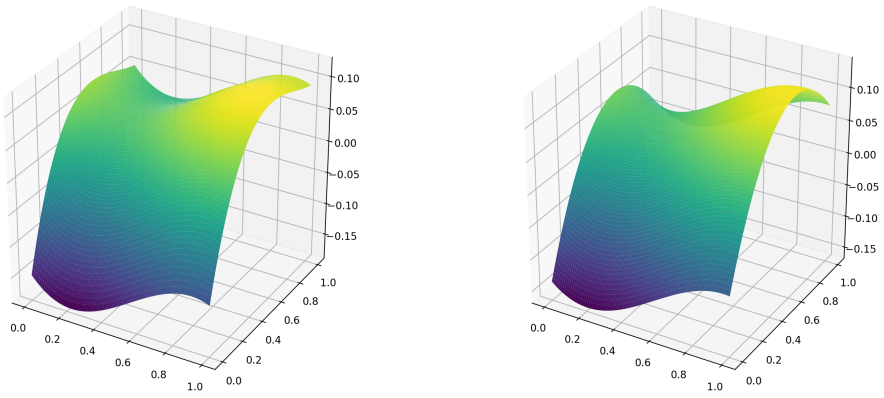


Figure 3: (Left) $g^\star_{1,1}(\mathbf{x})$ generated under additive model. (Right) $\widehat{g}_{1,1}(\mathbf{x})$ estimated under additive assumption.

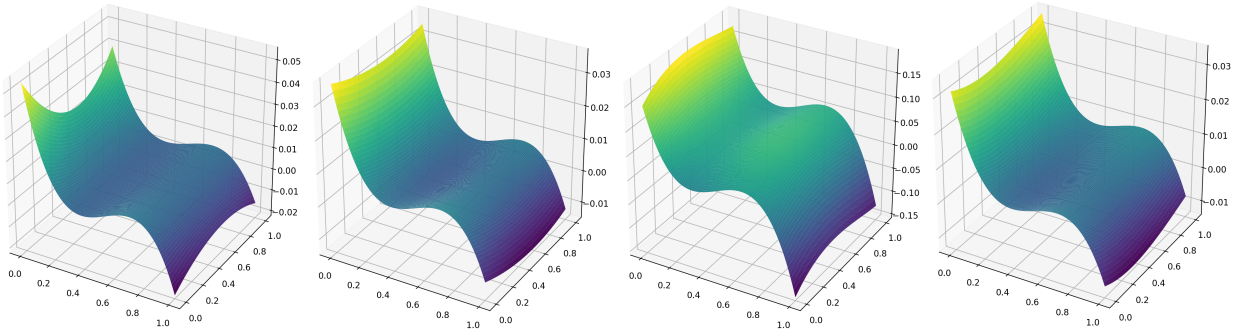
Further, we simulate the data such that the additive assumption (13) is not valid. Specifically, we generate $g_{m,r_m}(\mathbf{x}_{m,i_m\cdot})$ in a multiplicative scheme such that

$$g_{m,r_m}(\mathbf{x}_{m,i_m\cdot}) = \prod_{d_m=1}^{D_m} g_{m,r_m,d_m}(x_{m,i_m d_m}). \tag{33}$$

where g_{m,r_m,d_m} is given by (32). We conduct the IP-SVD procedure using the additive approximation (13). The true and estimated function of $g_{1,1}(\mathbf{x})$ are plotted in Figure 4a and 4c, respectively. The estimated function can capture some structures of the true function but misses other details as we approximate it with the additive form. Figure 4b depicts the projection of the true function $g^\star_{1,1}$ to the additive sieve space used in IP-SVD. The projection is supposed to be the best function estimate that can be obtained from the additive sieve basis. Note that $\mathbf{A}$ is identified up to an orthogonal matrix and so is $\mathbf{G}$. To address this potential problem of non-identifiability of $g_{1,1}$, we calculate the best linear combination of

$\widehat{g}_{1,r}$, $r = 1, \ldots, R$, that is closest to $g^*_{1,1}$ to mimic any potential orthogonal matrix applied to $\mathbf{G}$. The best linear combination is reported in Figure 4d. As one can see, Figure 4b and 4d are almost identical to each other. In conclusion, the projected Tucker under an additive basis assumption can ideally recover at most the linear (and additive) part of the true parametric component $\mathbf{G}_m$. The performance of this approximation depends on the deviation between the $\mathbf{G}_m$ and its projected version $\mathbf{P}_m\mathbf{G}_m$.

To assess the performance of IP-SVD when the additive assumption (13) becomes invalid, we repeat the experiment in Table 1 with exactly the same settings except that the multiplicative scheme in (33) is used to generate $\mathbf{G}_m$. The errors in estimating $\mathbf{A}_m$ and $\mathcal{Y}$ are reported in Table 5. Comparing Table 1 with Table 5, we observe that even when the additive assumption in (13) is not valid, IP-SVD still performs better than HOOI. But the improvement under misspecification is not as good as that under the valid additive assumption. This shows empirically that even when the additive assumption is violated, IP-SVD in general performances better than HOOI as long as the sieve basis used in IP-SVD can partially explain the parametric part of $\mathbf{A}_m$ with respect to $\mathbf{X}_m$.



(a) $g^\star_{1,1}(\mathbf{x})$ (b) $\mathbf{P}_1 g_{1,1}(\mathbf{x})$ (c) $\widehat{g}_{1,1}(\mathbf{x})$ (d) $\sum_{r=1}^{R} a_r \widehat{g}_{1,r}(\mathbf{x})$

Figure 4: (a) True function of $g^\star_{1,1}(\mathbf{x})$ (b) Projected version of $g^\star_{1,1}(\mathbf{x})$ (c) Estimated $g_{1,1}(\mathbf{x})$ (d) Best linear combination of $\widehat{g}_{1,r}(\mathbf{x})$.

Table 5: Under varying dimensions and signal-to-noise ratio. The mean and standard deviation of $\ell_2(\widehat{\mathbf{A}}_1)$ and $\text{ReMSE}_{\mathcal{Y}}$ from 100 replications when the additive loading assumption is replaced with the multiplicative assumption.

	α	0.1			0.3			0.5		
	I	100	200	300	100	200	300	100	200	300
IP-SVD	$\ell_2(\widehat{\mathbf{A}}_1)$	1.430 (0.103)	1.450 (0.126)	1.438 (0.115)	1.225 (0.165)	1.182 (0.210)	1.132 (0.203)	0.853 (0.212)	0.796 (0.217)	0.820 (0.228)
	$\text{ReMSE}_{\mathcal{Y}}$	2.596 (0.521)	2.395 (0.579)	2.375 (0.482)	1.189 (0.210)	1.095 (0.175)	0.984 (0.157)	0.741 (0.117)	0.709 (0.117)	0.704 (0.132)
HOOI	$\ell_2(\widehat{\mathbf{A}}_1)$	1.705 (0.012)	1.720 (0.006)	1.723 (0.005)	1.705 (0.012)	1.720 (0.005)	1.724 (0.004)	1.568 (0.213)	1.653 (0.168)	1.663 (0.188)
	$\text{ReMSE}_{\mathcal{Y}}$	8.092 (1.686)	10.108 (2.596)	12.049 (2.701)	3.263 (0.672)	3.874 (0.721)	3.911 (0.781)	1.528 (0.332)	1.538 (0.294)	1.513 (0.300)

6 Real data applications

6.1 Multi-variate Spatial-Temporal Data

In this section, we illustrate the usefulness of the STEFA model and the IP-SVD algorithm on the Comprehensive Climate Dataset (CCDS) – a collection of climate records of North America. The dataset was compiled from five federal agencies sources by Lozano et al. (2009)[3]. Specifically, we show that we can use the STEFA and IP-SVD to estimate interpretable loading functions, deal better with large noises and make more accurate predictions than the vanilla Tucker decomposition.

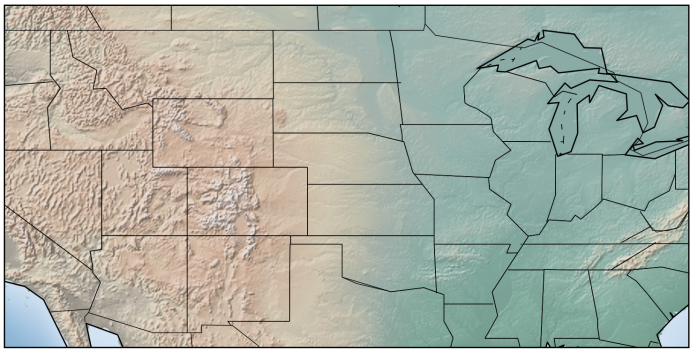


Figure 5: The geological region for which the data is collected

The data contains monthly observations of 17 climate variables from 1990 to 2001 on a

[3] http://www-bcf.usc.edu/~liu32/data/NA-1990-2002-Monthly.csv

2.5×2.5 degree grid for latitudes in $(30.475, 50.475)$, and longitudes in $(-119.75, -79.75)$. Figure 5 plots the geological region which covers the majority of the continent of United States and the southern part of Canada. The total number of observation locations is 125 and the whole time series spans from January, 1990 to December, 2001. Due to the data quality, we use only 16 measurements listed in Table 6 at each location and time point. Thus, the dimensions our our dataset are 125 (locations) $\times$ 16 (variables) $\times$ 156 (time points). Detailed information about data is given in Lozano et al. (2009).

Table 6: Variables and data sources in the Comprehensive Climate Dataset (CCDS)

Variables (Short name)	Variable group	Type	Source
Methane (CH4)	CH_4	Greenhouse Gases	NOAA
Carbon-Dioxide (CO2)	CO_2		
Hydrogen (H2)	H_2		
Carbon-Monoxide (CO)	CO		
Temperature (TMP)	TMP	Climate	CRU
Temp Min (TMN)	TMP		
Temp Max (TMX)	TMP		
Precipitation (PRE)	PRE		
Vapor (VAP)	VAP		
Cloud Cover (CLD)	CLD		
Wet Days (WET)	WET		
Frost Days (FRS)	FRS		
Global Horizontal (GLO)	SOL	Solar Radiation	NCDC
Direct Normal (DIR)	SOL		
Global Extraterrestrial (ETR)	SOL		
Direct Extraterrestrial (ETRN)	SOL		

We first focus on the spatial function structure of this data set. The covariates $\mathbf{X} \in \mathbb{R}^{125\times 2}$ of the spacial dimension contain the latitudes and longitudes of all sampling locations, which basically capture the spatial continuity of factor loadings on mode 1. The semi-parametric form (5) for this application is written as

$$\mathcal{Y} = \mathcal{F} \times_1 (\mathbf{\Phi}_1(\mathbf{X})\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}) + \mathbf{\Gamma}_1) \times_2 \mathbf{A}_2 \times_3 \mathbf{I} + \mathcal{E}. \tag{34}$$

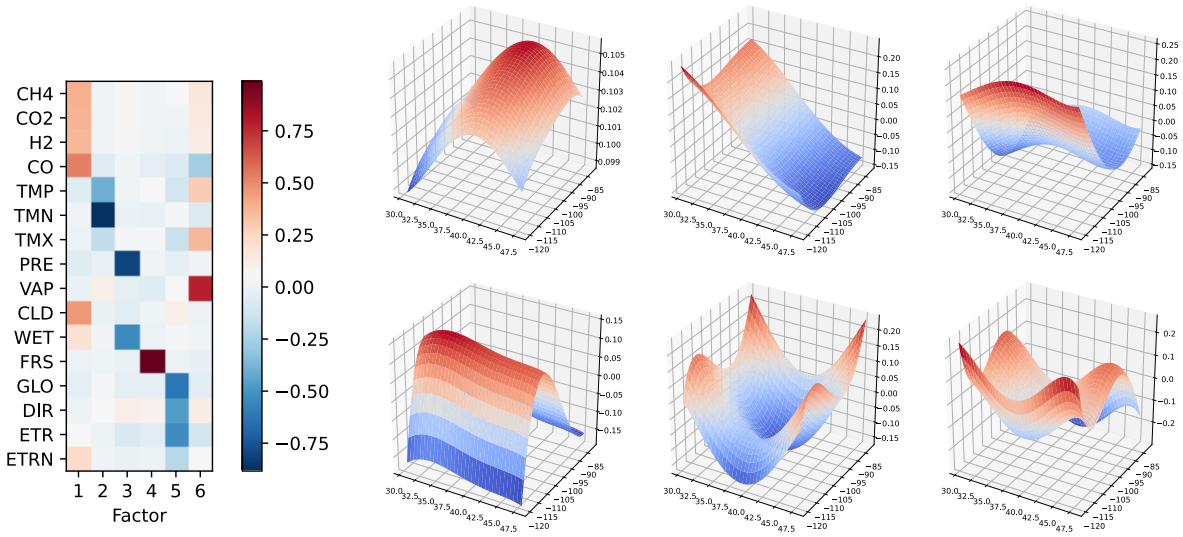
The first mode is the space dimension with loading matrix $\mathbf{A}_1 = \mathbf{\Phi}_1(\mathbf{X})\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}) + \mathbf{\Gamma}_1$. The second mode is the variable dimension with $\mathbf{A}_2$ as the variable loading matrix. The third mode is the time dimension which we do not compress. So we use the identity matrix $\mathbf{I}$ in place of $\mathbf{A}_3$. This is a matrix-variate factor model similar to Chen and Fan (2023) but

incorporates covariate effects on the loading matrix in the spatial-mode. We normalized each time series to have a unit ℓ_2 norm.

Climate variable and spatial factors. We use $R_1, R_2, R_3 = 6, 6, 156$ where the time mode is not compressed and the other two latent dimensions are chosen according to the literature (Lozano et al., 2009; Bahadori et al., 2014; Chen et al., 2020b). We use the Legendre basis functions of order 5 for $\mathbf{\Phi}_1(\mathbf{X})$ and number of basis $J = 11$. The slices of latent tensor factor $\mathbf{F}_{r_1::}$, $r_1 \in [6]$, correspond to six spatial factors and the slices $\mathbf{F}_{:r_2:}$, $r_2 \in [6]$, correspond to the six climate variable factors. The meaning of the latent factors can be inferred from their corresponding variable loading matrix $\mathbf{A}_2$ and spatial loading surfaces in $\mathbf{\Phi}_1(\mathbf{X})\mathbf{B}_1$.

Figure 6 (a) shows the heatmap of the varimax-rotated loading matrix $\mathbf{A}_2$. It is clear that the corresponding first climate factor weighted mostly on the four greenhouse gases. Thus, the first climate factor can be interpreted as the greenhouse gas factor. Interestingly, this greenhouse gas factor also loads heavily on cloud cover (CLD), echoing with a recent scientific research on the observational evidence between greenhouse gas and cloud covers (Ceppi and Nowack, 2021). In a similar way, the second to sixth climate variable factor can be interpreted as temperature, precipitation (wet), frost, solar, and vapor factors, respectively. The top six climate factors explain approximately 82.26% , 12.13%, 1.48%, 0.58%, 0.31%, 0.26% of the variance along the second (climate variable) mode of the tensor.

Figure 6 (b) presents six estimated bi-variate spatial loading surfaces corresponding to the six columns of $\mathbf{\Phi}_1(\mathbf{X})\widehat{\mathbf{B}}_1$. The space loading surfaces captures the common spatial variances in 16 environmental variables and they are highly nonlinear. More insights can be drawn by juxtaposing the discovered loading surfaces with the geological map in Figure 5 with aligned latitudes and longitudes. The high value (red) region in the first loading surface corresponds to the Great Lakes region of U.S. and Canada, which was highly-populated and has a well-developed industry in the 90's. The second surface represents a south-to-north gradient and a coast-to-inland gradient. The third surface has high values in the mountain region of U.S. The discovered top three major loading surfaces have their sociological and



(a) Variable Factors (b) Spatial Factors

Figure 6: (a) Heat map plots the varimax-rotated $\widehat{\mathbf{A}}_2$. The first six variable factors explain approximately 82.26% , 12.13%, 1.48%, 0.58%, 0.31%, 0.26% of the variance along the second mode of the tensor $\mathcal{Y}$. (b) Six surfaces are the estimated space loading surfaces plotted from six columns of $\widehat{\mathbf{G}}_1(\mathbf{X}) = \mathbf{\Phi}_1(\mathbf{X})\widehat{\mathbf{B}}_1$. From the top-left to the bottom right sub-figures correspond to the first to the sixth space loading functions with decreasing singular values. The coordinates of X and Y axis are aligned with the latitudes and longitudes in Figure 5.

geological correspondences. Beyond those, the estimates are very noisy and interpretation gets hard. The top sixth column of $\mathbf{\Phi}(\mathbf{X})_1\widehat{\mathbf{B}}_1$ explain approximately 93.16% , 2.39%, 0.75%, 0.42%, 0.18%, 0.14% of the variance along the first (spatial) mode of the tensor.

Fitting real data with different noise levels. In this section, we compare the vanilla and projected Tucker decomposition by their performances in fitting signal with different levels of noise. To generate different noise levels, we treat the estimated signal $\widehat{\mathcal{S}}_v$ and noise $\widehat{\mathcal{E}}_v$ from *vanilla* Tucker decomposition as the true signal $\mathcal{S}$ and noise $\mathcal{E}$ and calibrate the real data with different noise amplifier $\alpha > 0$. Specifically, the calibrated data is generated as $\mathcal{Y} = \widehat{\mathcal{S}}_v + \alpha \times \widehat{\mathcal{E}}_v$. The setting $\alpha = 1$ corresponds to the original data. We compare the relative mean square errors (ReMSE) of the signal estimator $\text{ReMSE}_{\mathcal{S}} = \|\mathcal{S} - \widehat{\mathcal{S}}\|_F^2 / \|\mathcal{S}\|_F^2$ for vanilla and projected Tucker decomposition in Figure 7. For the vanilla Tucker decomposition, we use the HOOI algorithm. For the projected Tucker decomposition, we use the same setting as previously, that is, we use the Legendre basis functions of order 5 for $\mathbf{\Phi}(\mathbf{X})$, number of basis $J = 11$ and latent dimensions $R_1, R_2, R_3 = 6, 6, 156$. Two methods behave the same in the noiseless case where $\alpha = 0$. However, in the noisy setting where $\alpha > 0$, the IP-SVD outperforms the HOOI at all noise levels.

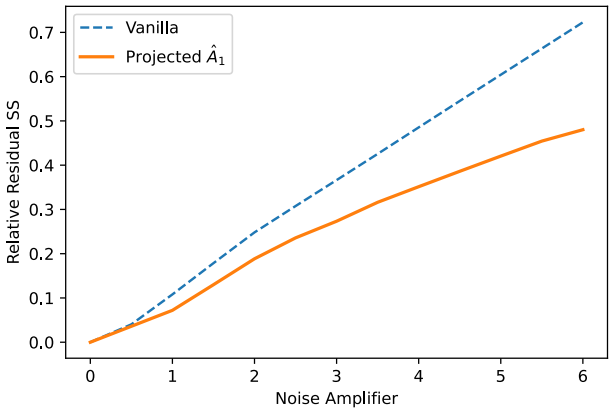


Figure 7: Relative mean square errors (ReMSE) by projected versus vanilla Tucker decomposition with different noise amplifiers. The relative residual SS of the signal part is defined as $\|\mathcal{S} - \widehat{\mathcal{S}}\|_F^2 / \|\mathcal{S}\|_F^2$ where $\widehat{\mathcal{S}} := \widehat{\mathcal{F}} \times_1 \widehat{\mathbf{A}}_1 \times_2 \widehat{\mathbf{A}}_2$. The loading $\mathbf{A}_1$, $\mathbf{A}_2$ and factor $\mathcal{F}$ are estimated by HOOI and IP-SVD, respectively, for vanilla and projected Tucker decomposition.

Spatial prediction. In this section, we compare the prediction performances of the methods based vanilla and projected Tucker decomposition. The two prediction procedures are presented in Section 3.4. We randomly choose the training set to be 50%, 67%, and 75% of the whole data set. Table 7 shows the prediction errors, average over cross validations, of the two methods respectively. It is clear that the STEFA model with projected Tucker decomposition outperforms the vanilla methods.

Table 7: Relative prediction error (averaged value by cross validation). For ease of display, the errors for Vanilla and Projected Tucker are reported as 100× the true value.

Training set proportion	50%	67%	75%
Vanilla	3.52	3.48	3.05
Projected	3.20	3.23	3.01
Improvement	9.0%	7.2%	1.3%

Temporal-mode compression. Now we consider fitting the real data with a more complex model where the mode corresponds to time is also compressed:

$$\mathcal{Y} = \mathcal{F} \times_1 (\mathbf{\Phi}_1(\mathbf{X})\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}) + \mathbf{\Gamma}_1) \times_2 \mathbf{A}_2 \times_3 (\mathbf{\Phi}_3(t)\mathbf{B}_3 + \mathbf{R}_3(t) + \mathbf{\Gamma}_3) + \mathcal{E}. \tag{35}$$

For the space mode, we use the same setting as previously, that is, we use the Legendre basis functions of order 5 for $\mathbf{\Phi}_1(\mathbf{X})$ and the number of basis $J_1 = 11$. For the time mode, we use the sinusoidal basis functions of order 12 for $\mathbf{\Phi}_3(t)$ and the number of basis $J_3 = 13$. Figure 8 presents the first two columns of $\mathbf{\Phi}_3(t)\mathbf{B}_3$ which explains approximately 80.69% and 0.14% of the variance along time mode of the tensor. Each column of the loading matrix $\mathbf{\Phi}_3(\mathbf{t})\widehat{\mathbf{B}}_3$ can be interpreted from its temporal pattern. The first time loading corresponds to the temporal mean since it is almost flat over time. The second time loading corresponds to a linear trend component. This trend coincides with the annual greenhouse gas emission data from U.S. environment protection agency[4], where the greenhouse gas emission has an overall increasing trend from 1992 to 2002 with local peaks around 1995 and 2000. The other time loading dimensions are less prominent as they account for a small portion of variations. As

[4]https://cfpub.epa.gov/ghgdata/inventoryexplorer

a result, their corresponding interpretations are not obvious and we omit their plots here.

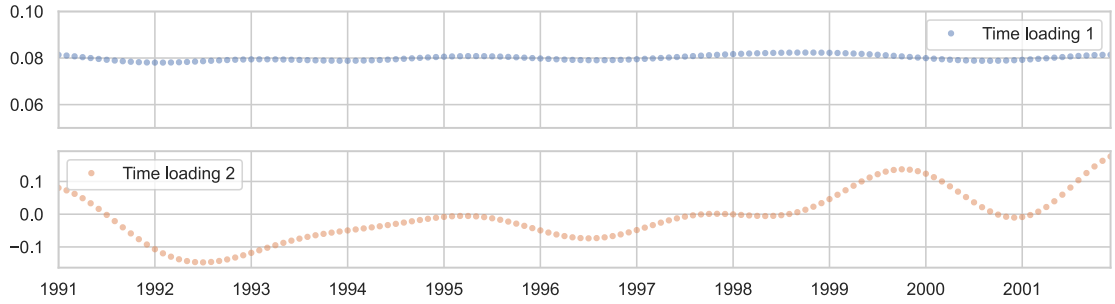


Figure 8: Top two functions of time that corresponds to the first two columns of $\mathbf{\Phi}_3(t)\mathbf{B}_3$.

6.2 Human Brain Connection Data

We illustrate another application of the STEFA model and IP-SVD to the human brain connection data (Desikan et al., 2006). This Human Connectome Project (HCP) dataset consists of brain structural networks collected from 136 individuals. Each brain network is represented as a 68×68 binary matrix, each entry of which encodes the presence or absence of a fiber connection between 68 brain regions. Thus, the final observation $\mathcal{Y}$ is of dimension $68 \times 68 \times 136$. Associated with 136 individuals, there are 573 features including ages, genders, and various measurements of their brains. This dataset has been used in Hu et al. (2022) for tensor regression and it is available in the R package **tensorregress**. We consider the instance of the STEFA model with $\mathbf{A}_3 = \mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3 + \mathbf{R}_3(\mathbf{X}) + \mathbf{\Gamma}_3$. The covariate $\mathbf{X}$ contains five features: gender (65 females vs. 71 males), age 22-25 ($n = 35$), age 26-30 ($n = 58$), and age 31+ ($n = 43$). These categorical variables are coded using sum-to-zero contrasts and the lower rank is set as $(10, 10, 4)$, the same as those set in Hu et al. (2022). As an illustration, we choose $\mathbf{\Phi}_3(\mathbf{X})$ generated by polynomial basis of order 1, which is a similar linear setting as that in Hu et al. (2022) with identity link function.

Hu et al. (2022) consider a similar setting for tensor regression. However, there are two major differences. First, they consider a generalized linear model (GLM) with a low-Tucker-rank coefficient tensor that predict $\mathcal{Y}$ with given observations $\mathbf{X}$ and their estimation is based on the maximum likelihood estimation. On the contrary, the STEFA model aims to discover

the relationships between entries in $\mathcal{Y}$ (with the help of covariates $\mathbf{X}$ when available) without using any distributional assumption. This relationship between entries in $\mathcal{Y}$ can also be used to predict values by interpolation as proposed in Section 3.4. Second, the STEFA model is comparable to MMC tensor regression only when the GLM linkage function is linear. In such setting, the form of the tensor regression model in Hu et al. (2022) is equivalent to assuming $\mathbf{A}_3 = \mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$ with $\mathbf{\Phi}_3(\mathbf{X}) = \mathbf{X}$ in the STEFA model. It has two limitations: only linear components of $\mathbf{X}$ is considered and the non-parametric residuals independent with $\mathbf{X}$ are ignored. In contrast, the STEFA model and the IP-SVD estimate $\mathbf{A}_3$ which consists of covariate-relevant $\mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$, the residual component $\mathbf{R}(\mathbf{X})$, as well as the orthogonal component $\mathbf{\Gamma}_3$. We compare the relative mean squared error $\text{ReMSE}_{\mathcal{Y}} = \left\|\widehat{\mathcal{Y}} - \mathcal{Y}\right\|_F^2 / \|\mathcal{Y}\|_F^2$ of the STEFA and MMC tensor regression with the identity link in Table 8. The relative mean squared error of the STEFA is smaller under all choices of basis functions. In fact, the STEFA model complements the work in Hu et al. (2022) since the orthogonal $\mathbf{\Gamma}_3$ can also be included in their tensor regression models.

Table 8: Relative mean squared error $\text{ReMSE}_{\mathcal{Y}} = \left\|\widehat{\mathcal{Y}} - \mathcal{Y}\right\|_F^2 / \|\mathcal{Y}\|_F^2$ of the MMC with identity link and the STEFA. Each columns corresponds to a type of basis function with its order in the parentheses. The MMC and the STEFA use the same basis function in $\mathbf{\Phi}(\mathbf{X})$.

	Polynomials (1)	Polynomials (3)	Legendre (5)
MMC (Identity link)	39.6%	39.5%	39.4%
STEFA	38.7%	38.4%	38.3%
Improvement	2.3%	2.8%	2.8%

The covariate-relevant $\mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$ is determined by covariates specified by domain experts. Researchers may be curious to identify features affecting brain connectivity other than those already known in the field. Now we show that the residual $\widehat{\mathbf{\Gamma}}_3 = \widehat{\mathbf{A}}_3 - \widehat{\mathbf{\Phi}}_3(\mathbf{X})\widehat{\mathbf{B}}_3$ obtained by STEFA can be used to discover several features other than gender and age. Analogous to the interpretation that the rows of the 136×4 loading matrix $\widehat{\mathbf{A}}_3$ can be reviewed as the low-rank representation of 136 subjects in the latent factor space, matrices $\widehat{\mathbf{\Gamma}}_3$ and $\widehat{\mathbf{\Phi}}_3(\mathbf{X})\widehat{\mathbf{B}}_3$ can be interpreted, respectively, as covariate-independent and covariate-dependent low-rank representations. The idea to identify important features left in the residual component

is to first use $\widehat{\mathbf{\Gamma}}_3$ in spectral clustering to divide 136 subjects into four groups, and then use recursive feature elimination (RFE) to identify the top four important features that differentiate these four groups. The identified top four important features that are disparate across groups are the average thickness of the right transverse temporal gyri which has been shown to be correlated with human acoustic processing (Warrier et al., 2009), the volumn of accumbens area which is a key structure in mediating emotional and motivation processing, modulating reward and pleasure processing, and serving a key limbic-motor interface (Cohen et al., 2009; Salgado and Kaplitt, 2015), the unadjusted negative emotion affect related to sadness, fear, and anger, and a personality raw score on being active or not. Section **??** provides another illustration of using the STEFA and IP-SVD for explanatory data analysis to partitioning the brain connectivity according to the covariate-relevant loading $\mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$. These interesting discoveries from explanatory data analysis can be used as good starting points for the following more rigorous scientific researches.

7 Discussion

This paper introduces a high-dimensional Semiparametric TEnsor FActor (STEFA) model with nonparametric loading functions that depend on a few observed covariates. This model is motivated by the fact that observed variables can partially explain the factor loadings, which helps to increase the accuracy of estimation and the interpretability of results. We propose a computationally efficient algorithm IP-SVD to estimate the unknown tensor factor, loadings, and the latent dimensions. The advantages of IP-SVD are two-fold. First, unlike HOOI which iterates in the ambient dimension, IP-SVD finds the principal components in the covariate-related subspace whose dimension can be significantly smaller. As a result, IP-SVD requires weaker SNR conditions for convergence. Secondly, the projection also reduces the effect dimension size of stochastic noise and thus IP-SVD yields an estimate of latent factors with faster convergence rates.

While tensor data is everywhere in the physical world, statistical analysis for tensor data is still challenging. There are several interesting topics for future research. First, it is important

to develop non-parametric tests on whether observed relevant covariates have explaining powers on the loadings and whether they fully explain the loadings. However, under the tensor decomposition setting, this is more challenging than a straightforward extension from Fan et al. (2016). Second, we mentioned briefly that, when there are multiple observations, one can apply IP-SVD on the sample covariance tensor. However, a more precise algorithm is needed. Last but not the least, it is of great need to develop new methods to use STEFA in tensor regression or other tensor data related applications.

Acknowledgments

We express our gratitude to the referees and the editors for their invaluable feedback, which greatly enhanced the quality of this paper.
Conflict of interest: None declared.

Funding

Chen's research is supported in part by NSF Grant DMS-1803241. Xia's research is supported in part by Hong Kong RGC ECS Grant 26302019 and GRF grant 16303320. Fan's research was partially supported by NSF Grants DMS-1662139, DMS-1712591, DMS-2210833, and ONR grant N00014-22-1-2340.

Data, Source Code, and Supplementary Material

The Comprehensive Climate Dataset (CCDS) is available at `http://www-bcf.usc.edu/~liu32/data/NA-1990-2002-Monthly.csv`. The source code is available at `https://github.com/ElynnCC/STEFA-Code`. Supplementary material is available online at *Journal of the Royal Statistical Society: Series B.*

References

Evrim Acar, Tamara G Kolda, and Daniel M Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv preprint arXiv:1105.3422*, 2011.

Seung C Ahn and Alex R Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

Genevera Allen. Sparse higher-order principal components analysis. In *Artificial Intelligence and Statistics*, pages 27–36, 2012a.

Genevera I Allen. Regularized tensor factorizations and higher-order principal components analysis. *arXiv preprint arXiv:1202.2476*, 2012b.

Theodore W Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150, 1956.

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*. 1962.

Mohammad Taha Bahadori, Qi Rose Yu, and Yan Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in neural information processing systems*, pages 3491–3499, 2014.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai, Kunpeng Li, et al. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465, 2012.

Muthu Baskaran, Tom Henretty, Benoit Pradelle, M Harper Langston, David Bruns-Smith, James Ezick, and Richard Lethin. Memory-efficient parallel tensor decompositions. In

2017 IEEE High Performance Extreme Computing Conference (HPEC), pages 1–7. IEEE, 2017.

Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, pages 1861–1872, 2019.

Raymond J Carroll and David Ruppert. Discussion: Conditional growth charts. *The Annals of Statistics*, 34(5):2098–2104, 2006.

Paulo Ceppi and Peer Nowack. Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences*, 118(30):e2026290118, 2021.

Elynn Y. Chen and Jianqing Fan. Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 118(542):1038–1055, 2023.

Elynn Y. Chen, Ruey S. Tsay, and Rong Chen. Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, pages 1–37, 2019.

Elynn Y. Chen, Dong Xia, Chencheng Cai, and Jianqing Fan. Supplement to "semiparametric tensor factor analysis by iteratively projected svd". 2020a.

Elynn Y. Chen, Xin Yun, Qiwei Yao, and Rong Chen. Modeling multivariate spatial-temporal data with latent low-dimensional dynamics. *arXiv preprint arXiv:2002.01305*, 2020b.

Rong Chen, Dan Yang, and Cun-Hui Zhang. Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116, 2022.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Michael X Cohen, Nikolai Axmacher, Doris Lenartz, Christian E Elger, Volker Sturm, and Thomas E Schlaepfer. Good vibrations: cross-frequency coupling in the human nucleus

accumbens during reward processing. *Journal of cognitive neuroscience*, 21(5):875–889, 2009.

Gregory Connor and Oliver Linton. Semiparametric estimation of a characteristic-based factor model of common stock returns. *Journal of Empirical Finance*, 14(5):694–717, 2007.

Gregory Connor, Matthias Hagmann, and Oliver Linton. Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica*, 80(2):713–754, 2012.

André LF De Almeida and Alain Y Kibangou. Distributed large-scale tensor decomposition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 26–30. IEEE, 2014.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000a.

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, 2000b.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Jianqing Fan, Yuan Liao, and Weichen Wang. Projected principal component analysis in factor models. *Annals of statistics*, 44(1):219, 2016.

Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical Foundations of Data Science*. Chapman & Hall/CRC, 2020.

Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, and Ziwei Zhu. Robust high dimensional factor models with applications to statistical machine learning. *Statistics Science*, art. arXiv:1808.03889, 2020.

Yuefeng Han, Rong Chen, Dan Yang, and Cun-Hui Zhang. Tensor factor model estimation by iterative projection. *arXiv preprint arXiv:2006.02611*, 2020.

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

Botao Hao, Boxiang Wang, Pengyuan Wang, Jingfei Zhang, Jian Yang, and Will Wei Sun. Sparse tensor additive regression. *The Journal of Machine Learning Research*, 22(1):2989–3031, 2021.

Jiaxin Hu, Chanwoo Lee, and Miaoyan Wang. Generalized tensor decomposition with features on multiple modes. *Journal of Computational and Graphical Statistics*, 31(1):204–218, 2022.

Jianhua Z Huang, Haipeng Shen, and Andreas Buja. The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association*, 104(488):1609–1620, 2009.

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

Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, pages 694–726, 2012.

Gen Li, Dan Yang, Andrew B Nobel, and Haipeng Shen. Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, 146:7–17, 2016.

Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change

attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596. ACM, 2009.

Xiaojun Mao, Song Xi Chen, and Raymond KW Wong. Matrix completion with covariate information. *Journal of the American Statistical Association*, 114(525):198–210, 2019.

Garvesh Raskutti, Ming Yuan, Han Chen, et al. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

Sanjay Salgado and Michael G Kaplitt. The nucleus accumbens: a comprehensive review. *Stereotactic and functional neurosurgery*, 93(2):75–93, 2015.

Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

Bernard W Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996.

Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu. Tensor completion algorithms in big data analytics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):1–48, 2019.

Will Wei Sun and Lexin Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Dong Wang, Xialu Liu, and Rong Chen. Factor models for matrix-valued high-dimensional time series. *Journal of econometrics*, 208(1):231–248, 2019.

Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *The Journal of Machine Learning Research*, 21(1):6146–6183, 2020.

Miaoyan Wang and Yun Song. Tensor decompositions via two-mode higher-order svd (hosvd). In *Artificial Intelligence and Statistics*, pages 614–622, 2017.

Wen-Ting Wang and Hsin-Cheng Huang. Regularized principal component analysis for spatial data. *Journal of Computational and Graphical Statistics*, 26(1):14–25, 2017.

Catherine Warrier, Patrick Wong, Virginia Penhune, Robert Zatorre, Todd Parrish, Daniel Abrams, and Nina Kraus. Relating structure to function: Heschl's gyrus and acoustic processing. *Journal of Neuroscience*, 29(1):61–69, 2009.

Dong Xia and Fan Zhou. The sup-norm perturbation of hosvd and low rank tensor denoising. *Journal of Machine Learning Research*, 20(61):1–42, 2019.

Zhuoyan Xu, Jiaxin Hu, and Miaoyan Wang. Generalized tensor regression with covariates on multiple modes. *arXiv preprint arXiv:1910.09499*, 2019.

Anru Zhang. Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47(2): 936–964, 2019.

Anru Zhang and Rungang Han. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, 0(0):1–34, 2019.

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

Jie Zhou, Will Wei Sun, Jingfei Zhang, and Lexin Li. Partially observed dynamic tensor response regression. *Journal of the American Statistical Association*, pages 1–16, 2021.

Lan Zhou and Huijun Pan. Principal component analysis of two-dimensional functional data. *Journal of Computational and Graphical Statistics*, 23(3):779–801, 2014.

List of Figures

Supplementary Material of "Semiparametric Tensor Factor Analysis by Iteratively Projected SVD"

Elynn Y. Chen[1], Dong Xia[2], Chencheng Cai[3] and Jianqing Fan[4,5]

[1]New York University, [2]Hong Kong University of Science and Technology

[3]Washington State University, [4] Fudan University, [5]Princeton University

arXiv:2007.02404v2 [stat.ME] 2 Apr 2024

Appendix A Major Theoretical Proofs

Proof of Lemma **??**. Let $\mathcal{S} = \widetilde{\mathcal{F}} \times_1 \widetilde{\mathbf{A}}_1 \times_2 \cdots \times_M \widetilde{\mathbf{A}}_M$ be a Tucker decomposition of $\mathcal{S}$ such that $\widetilde{\mathcal{F}} \in \mathbb{R}^{R_1 \times \cdots \times R_M}$ is the core tensor, and for $m \in [M]$, $\widetilde{\mathbf{A}}_m \in \mathbb{R}^{I_m \times R_m}$ is the m-mode loading matrix satisfying Assumption **??**(i).

Clearly, $\mathcal{S} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \cdots \times_M \mathbf{A}_M$ is a valid Tucker decomposition satisfying Assumption **??**(i) if and only if there exist orthogonal matrices $\mathbf{H}_m \in \mathbb{O}^{R_m \times R_m}$ for all $m \in [M]$ such that $\mathbf{A}_m = \widetilde{\mathbf{A}}_m \mathbf{H}_m$ and $\mathcal{F} = \widetilde{\mathcal{F}} \times_1 \mathbf{H}_1^\top \times_2 \cdots \times_m \mathbf{H}_M^\top$. Moreover, then $\mathcal{M}_m(\mathcal{S})\mathcal{M}_m(\mathcal{S})^\top / I_m$ has the same singular values with $\mathcal{M}_m(\mathcal{F})\mathcal{M}_m(\mathcal{F})^\top$

Now we are in the place to prove the lemma. Recall that any Tucker decomposition of $\mathcal{S}$ satisfying Assumption **??**(i) are indexed by the set of orthogonal matrices $(\mathbf{H}_1, \cdots, \mathbf{H}_M)$ in reference to the decomposition $\mathcal{S} = \widetilde{\mathcal{F}} \times_1 \widetilde{\mathbf{A}}_1 \times_2 \cdots \times_M \widetilde{\mathbf{A}}_M$. The corresponding core tensor $\mathcal{F}$ is $\mathcal{F} = \widetilde{\mathcal{F}} \times_1 \mathbf{H}_1^\top \times_2 \cdots \times_M \mathbf{H}_M^\top$. The mode-$m$ matricization of $\mathcal{F}$ is $\mathcal{M}_m(\mathcal{F}) = \mathbf{H}_m^\top \mathcal{M}_m(\widetilde{\mathcal{F}}) \left[\bigotimes_{l \in [M], l \neq m} \mathbf{H}_l^\top\right]$. Suppose for some $\mathbf{H}_m$, Assumption **??**(i) is satisfied such that $\mathcal{M}_m(\mathcal{F})\mathcal{M}_m(\mathcal{F})^\top = \mathbf{H}_m^\top \mathcal{M}_m(\widetilde{\mathcal{F}})\mathcal{M}_m(\widetilde{\mathcal{F}})^\top \mathbf{H}_m = \mathbf{D}_m$ for some diagonal matrix $\mathbf{D}_m$ with non-zero decreasing diagonal entries. Then the diagonal entries of $\mathbf{D}_m$ are the eigenvalues of $\mathcal{M}_m(\widetilde{\mathcal{F}})\mathcal{M}_m(\widetilde{\mathcal{F}})^\top$ and the columns in $\mathbf{H}_m$ are the corresponding eigenvectors, because of the equality $\mathcal{M}_m(\widetilde{\mathcal{F}})\mathcal{M}_m(\widetilde{\mathcal{F}})^\top \mathbf{H}_m = \mathbf{H}_m \mathbf{D}_m$. Note that $\mathbf{D}_m$ has the same singular values with $\mathcal{M}_m(\mathcal{S})\mathcal{M}_m(\mathcal{S})^\top$. As a result, when the singular values of $\mathcal{M}_m(\mathcal{S})\mathcal{M}_m(\mathcal{S})^\top$ are distinct, the eigenvalues and eigenvectors of $\mathcal{M}_m(\widetilde{\mathcal{F}})\mathcal{M}_m(\widetilde{\mathcal{F}})^\top$ can be uniquely identified (up to a global sign), resulting in an unique $\mathbf{H}_m$. Here, the uniqueness is up to a column-wise sign of $\mathbf{H}_m$.

In conclusion, starting from an arbitrary Tucker decomposition $\mathcal{S} = \widetilde{\mathcal{F}} \times_1 \widetilde{\mathbf{A}}_1 \times_2 \cdots \times_M \widetilde{\mathbf{A}}_M$ satisfying Assumption **??**(i), by choosing the columns of $\mathbf{H}_m$ to be the eigenvectors of

$\mathcal{M}_m(\widetilde{\mathcal{F}})\mathcal{M}_m(\widetilde{\mathcal{F}})^\top$ in descending order of eigenvalues, the Tucker decomposition with $\mathcal{F} = \widetilde{\mathcal{F}} \times_1 \mathbf{H}_1^\top \times_2 \cdots \times_M \mathbf{H}_M^\top, \mathbf{A}_1 = \widetilde{\mathbf{A}}_1\mathbf{H}_1, \cdots, \mathbf{A}_M = \widetilde{\mathbf{A}}_M\mathbf{H}_M$ is the unique Tucker decomposition satisfying both Assumptions 1(i) and 1(ii) when the eigenvalues of $\mathcal{M}_m(\mathcal{S})\mathcal{M}_m(\mathcal{S})^\top$ are distinct for all $m \in [M]$. □

Proof of Lemma **??**. We prove the initialization error and convergence of IP-SVD separately. Without loss of generality, we assume $\mathbb{E}\varepsilon_\omega^2 = 1$ for all $\omega \in [I_1] \times [I_2] \times [I_3]$.

We begin with the upper bound of the remainder term $\mathbf{R}_m(\mathbf{X}_m)$. By Assumption **??**, for each function $m \in [M]$ and $1 \le r_m \le R_m$, $|g_{m,r_m}(\mathbf{x}_{i_m}) - \mathbf{b}_{m,r_m}^\top \boldsymbol{\phi}_m(\mathbf{x}_{i_m})| = O(J_m^{-\tau/2})$ which bounds the (i_m, r_m)-th entry of $\mathbf{R}_m(\mathbf{X}_m)$. Therefore, a simple fact is

$$\|\mathbf{R}_m(\mathbf{X}_m)\|_{\mathrm{F}}^2/I_m = O\big(R_m \cdot J_m^{-\tau}\big) \tag{1}$$

for all $m \in [M]$.

Initialization error. Without loss of generality, we only prove the upper bound of $\|\widetilde{\mathbf{G}}_1^{(0)}\widetilde{\mathbf{G}}_1^{(0)\top} - \mathbf{G}_1\mathbf{G}_1^\top\|_{\mathrm{F}}$. Recall that $\mathbf{G}_1 = \boldsymbol{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1 + \mathbf{R}_1(\mathbf{X}_1)$ where, by the definition of $\boldsymbol{\Phi}_1(\mathbf{X}_1)$, we have $\boldsymbol{\Phi}_1(\mathbf{X}_1)^\top\boldsymbol{\Gamma}_1 = \mathbf{0}$. Let $\check{\mathbf{G}}_1/\sqrt{I_1}$ denotes the top-R_1 left singular vectors of $\boldsymbol{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1$. By Condition (1) and Davis-Kahan theorem (**?**) or spectral perturbation formula (**?**),

$$\|\check{\mathbf{G}}_1\check{\mathbf{G}}_1^\top - \mathbf{G}_1\mathbf{G}_1^\top\|_{\mathrm{F}}/I_1 = O\big(\sqrt{R_1} \cdot J_1^{-\tau/2}\big) \tag{2}$$

where we used the fact that $\sigma_{R_1}\big(\boldsymbol{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1/\sqrt{I_1}\big) \ge 1 - O(\sqrt{R_1} \cdot J_1^{-\tau/2}) \ge 1/2$ and $\|\mathbf{R}_1(\mathbf{X}_1)\|_{\mathrm{F}}/\sqrt{I_1} = O(\sqrt{R_1} \cdot J_1^{-\tau/2})$.

Recall that $\widetilde{\mathcal{Y}} = \mathcal{F} \times_1 (\mathbf{P}_1\mathbf{G}_1) \times_2 (\mathbf{P}_2\mathbf{G}_2) \times_3 (\mathbf{P}_3\mathbf{G}_3) + \mathcal{E} \times_1 \mathbf{P}_1 \times_2 \mathbf{P}_2 \times_3 \mathbf{P}_3$ and as a result

$$\mathcal{M}_1(\widetilde{\mathcal{Y}}) = \mathbf{P}_1\mathbf{G}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2) \otimes (\mathbf{P}_3\mathbf{G}_3)\big)^\top + \mathbf{P}_1\mathcal{M}_1(\mathcal{E})(\mathbf{P}_2 \otimes \mathbf{P}_3)^\top$$

where we denote $\otimes$ the kronecker product. The projector matrix $\mathbf{P}_m \in \mathbb{R}^{I_m \times I_m}$ with $\mathrm{rank}(\mathbf{P}_m) = J_m$. Denote the eigen-decomposition of $\mathbf{P}_m$ by $\mathbf{P}_m = \mathbf{U}_m\mathbf{U}_m^\top$ where $\mathbf{U}_m^\top\mathbf{U}_m =$

$\mathbf{I}_{J_m}$. Therefore, we write

$$\mathbf{P}_1\mathcal{M}_1(\mathcal{E})(\mathbf{P}_2\otimes\mathbf{P}_3)=\mathbf{U}_1\big(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\big)(\mathbf{U}_2\otimes\mathbf{U}_3)^\top.$$

In addition, we write

$$\begin{aligned}\mathbf{P}_1\mathbf{G}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top=&\,\mathbf{P}_1\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top\\&+\mathbf{P}_1\mathbf{R}_1(\mathbf{X}_1)\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top\end{aligned}$$

where the left singular space of the first matrix is the same to column space of $\check{\mathbf{G}}_1$. Denote the top-R_1 left singular vectors of $\mathbf{P}_1\mathbf{G}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top$ by $\mathring{\mathbf{G}}_1/\sqrt{I_1}$. Again by Davis-Kahan theorem (**?**) or spectral perturbation formula (**?**), we have

$$\|\mathring{\mathbf{G}}_1\mathring{\mathbf{G}}_1^\top-\check{\mathbf{G}}_1\check{\mathbf{G}}_1^\top\|_{\mathrm{F}}/I_1=O(\kappa_0\sqrt{R_1}\cdot J_1^{-\tau/2}) \tag{3}$$

where κ_0 is $\mathcal{F}$'s condition number and we used the facts

$$\begin{aligned}\sigma_{R_1}\Big(\mathbf{P}_1\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top\Big)\geq&\lambda_{\min}\sqrt{I_2I_3}\cdot\sigma_{R_1}\big(\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1\big)\\\geq&\,\lambda_{\min}\sqrt{I_1I_2I_3}/2\end{aligned}$$

and

$$\begin{aligned}\Big\|\mathbf{P}_1\mathbf{R}_1(\mathbf{X}_1)\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top\Big\|_{\mathrm{F}}=&O\big(\kappa_0\lambda_{\min}\sqrt{I_2I_3}\cdot\|\mathbf{R}_1(\mathbf{X}_1)\|_{\mathrm{F}}\big)\\=&\,O(\kappa_0\lambda_{\min}\sqrt{R_1I_1I_2I_3}\cdot J_1^{-\tau/2}).\end{aligned}$$

For notational simplicity, we write $\mathcal{M}_1(\widetilde{\mathcal{Y}})=\mathbf{A}_1+\mathbf{Z}_1$ where $\mathbf{A}_1=\mathbf{P}_1\mathbf{G}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top$ and $\mathbf{Z}_1=\mathbf{P}_1\mathcal{M}_1(\mathcal{E})(\mathbf{P}_2\otimes\mathbf{P}_3)^\top$. Then, $\mathring{\mathbf{G}}_1/\sqrt{I_1}$ are the top-R_1 left singular vectors of $\mathbf{A}_1$ and are also the top-R_1 eigenvectors of $\mathbf{A}_1\mathbf{A}_1^\top$. Since $\widetilde{\mathbf{G}}_1^{(0)}/\sqrt{I_1}$ are the top-R_1 left singular vectors of $\mathcal{M}_1(\widetilde{\mathcal{Y}})$, they are also the top-$R_1$ eigenvectors of $\mathcal{M}_1(\widetilde{\mathcal{Y}})\mathcal{M}_1^\top(\widetilde{\mathcal{Y}})$ whcih can be written as $\mathcal{M}_1(\widetilde{\mathcal{Y}})\mathcal{M}_1^\top(\widetilde{\mathcal{Y}})=\mathbf{A}_1\mathbf{A}_1^\top+\mathbf{A}_1\mathbf{Z}_1^\top+\mathbf{Z}_1\mathbf{A}_1^\top+\mathbf{Z}_1\mathbf{Z}_1^\top$. Observe that

$$\mathbf{Z}_1\mathbf{Z}_1^\top=\mathbf{U}_1\big(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\big)\big(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\big)^\top\mathbf{U}_1^\top.$$

Then, we can write

$$\mathcal{M}_1(\widetilde{\mathcal{Y}})\mathcal{M}_1^\top(\widetilde{\mathcal{Y}}) = \mathbf{A}_1\mathbf{A}_1^\top + J_2J_3\mathbf{P}_1 + \mathbf{A}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{A}_1^\top + \mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1.$$

By definition, the column space of $\mathring{\mathbf{G}}_1$ is a subspace of the column space of $\mathbf{P}_1$, and $\mathcal{P}_{\mathring{\mathbf{G}}_1}\mathbf{A}_1\mathbf{A}_1^\top\mathcal{P}_{\mathring{\mathbf{G}}_1} = \mathbf{A}_1\mathbf{A}_1^\top$ where $\mathcal{P}_{\mathring{\mathbf{G}}_1} = \mathring{\mathbf{G}}_1\mathring{\mathbf{G}}_1^\top/I_1$ denotes the orthogonal projection onto the column space of $\mathring{\mathbf{G}}_1$.

We write

$$\mathcal{M}_1(\widetilde{\mathcal{Y}})\mathcal{M}_1^\top(\widetilde{\mathcal{Y}}) = \underbrace{\big(\mathbf{A}_1\mathbf{A}_1 + J_2J_3\mathbf{P}_1\big)}_{\mathbf{M}} + (\mathbf{A}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{A}_1^\top) + \mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1. \tag{4}$$

Clearly, the top-R_1 left singular space of $\mathbf{M}$ is the column space of $\mathring{\mathbf{G}}_1$ and $\sigma_{R_1}(\mathbf{M}) - \sigma_{R_1+1}(\mathbf{M}) = \sigma_{R_1}(\mathbf{A}_1\mathbf{A}_1^\top) \geq \lambda_{\min}^2 \cdot I_1I_2I_3$.

The upper bounds on the spectral norm $\|\mathbf{A}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{A}_1^\top\|$ and $\|\mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1\|$ are due to the following lemma whose proof is postponed to Appendix E.

Lemma 1. *Suppose that Assumption* **??** *holds. There exists an absolute constant* $C_1 > 0$ *such that with probability at least* $1 - 2I_1^{-2}$,

$$\|\mathbf{A}_1\mathbf{Z}_1^\top\| \leq C_4\kappa_0\lambda_{\min}(I_1I_2I_3)^{1/2} \cdot \sqrt{J_1}\log^2 I_1.$$

By Lemma 1, the following bound holds with probability at least $1 - 2I_1^{-2}$

$$\|\mathbf{A}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{A}_1^\top\| \leq C_4\kappa_0\lambda_{\min}(I_1I_2I_3)^{1/2} \cdot \sqrt{J_1}\log^2 I_1.$$

Observe that

$$\begin{aligned}&\mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1\\ &\quad= \mathbf{U}_1\Big(\big(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\big)\big(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\big)^\top - J_2J_3\mathbf{I}_{J_1}\Big)\mathbf{U}_1^\top.\end{aligned}$$

Lemma 2. *Suppose that Assumption* **??** *holds. There exist absolute constants* $C_3, C_4 > 0$ *so*

that

$$\mathbb{P}\big(\|\mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1\| \geq C_3\sqrt{J_1J_2J_3}\log^4 I_1 + C_4J_1\log^{5/2} I_1\big) \leq 5I_1^{-2}.$$

By Lemma 2, with probability at least $1-5I_1^{-2}$,

$$\big\|\mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1\big\| \leq C_3\sqrt{J_1J_2J_3}\log^4 I_1 + C_4J_1\log^{5/2} I_1.$$

We now continue from (4). By Davis-Kahan theorem (**?**) and Lemma 1 and 2, we get with probability at least $1-7I_1^{-2}$ that

$$\big\|\mathring{\mathbf{G}}_1\mathring{\mathbf{G}}_1^\top - \widetilde{\mathbf{G}}_1^{(0)}\widetilde{\mathbf{G}}_1^{(0)\top}\big\|_{\mathrm{F}} \leq C_3\frac{\kappa_0\sqrt{R_1J_1}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}} \quad + C_4R_1^{1/2}\frac{(J_1J_2J_3)^{1/2}\log^4 I_1 + J_1\log^{5/2} I_1}{\lambda_{\min}^2 I_1I_2I_3}$$

for some absolute constants $C_3, C_4 > 0$. Denote the above event by $\mathfrak{E}_1$. Together with (2) and (3), we get on event $\mathfrak{E}_1$ that

$$\begin{aligned}&\|\mathbf{G}_1\mathbf{G}_1^\top - \widetilde{\mathbf{G}}_1^{(0)}\widetilde{\mathbf{G}}_1^{(0)\top}\|_{\mathrm{F}}/I_1\\ &\quad\leq C_3'\left(\frac{\kappa_0\sqrt{R_1J_1}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}} + \frac{\sqrt{R_1J_1(J_1\vee J_2J_3)}\log^4 I_1}{\lambda_{\min}^2 I_1I_2I_3} + \kappa_0\sqrt{R_1}\cdot J_1^{-\tau/2}\right). \qquad (5)\end{aligned}$$

In view of (5), the initialization $\widetilde{\mathbf{G}}_1^{(0)}$ is close to the truth as long as

$$\lambda_{\min}\sqrt{I_1I_2I_3} \geq C_1'\left(\kappa_0\sqrt{R_1J_1}\log^2 I_1 + \big(R_1J_1(J_1\vee J_2J_3)\big)^{1/4}\log^2 I_1\right)$$

and $J_1 \gg \kappa_0^{2/\tau}R_1^{1/\tau}$. The proof is completed by assuming $\kappa_0 = O(1)$ and $J_1 \asymp J_2 \asymp J_3$. Assuming warm initializations and iterates for $\widetilde{\mathbf{G}}_m^{(t-1)}$, we prove the contraction property for $\widetilde{\mathbf{G}}_m^{(t)}$.

IP-SVD iterations. Without loss of generality, we fix an integer value of t and prove the contraction inequality (**??**) for $m=1$. For notation simplicity, we denote

$$\mathrm{Err}_t = \max_{m=1,2,3}\|\widetilde{\mathbf{G}}_m^{(t)}\widetilde{\mathbf{G}}_m^{(t)\top} - \mathbf{G}_m\mathbf{G}_m^\top\|_{\mathrm{F}}/I_m.$$

By projected power iteration in Section **??**, the scaled singular vectors $\widetilde{\mathbf{G}}_1^{(t)}$ is obtained

by

$$\widetilde{\mathbf{G}}_1^{(t)}/\sqrt{I_1} = \mathrm{SVD}_{R_1}\Big(\mathbf{P}_1\mathcal{M}_1\Big(\mathcal{Y}\times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\Big)\Big)$$

Recall $\mathbf{A}_m = \mathbf{G}_m + \mathbf{\Gamma}_m$ for $m = 1, 2, 3$ and

$$\mathcal{Y} = \underbrace{\mathcal{F}\times_1 (\mathbf{G}_1+\mathbf{\Gamma}_1)\times_2(\mathbf{G}_2+\mathbf{\Gamma}_2)\times_3(\mathbf{G}_3+\mathbf{\Gamma}_3)}_{\mathcal{S}} + \mathcal{E}.$$

We then write

$$\begin{aligned}&\mathbf{P}_1\mathcal{M}_1\Big(\mathcal{Y}\times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\Big)\\ &\quad=\mathbf{P}_1\mathcal{M}_1\Big(\mathcal{S}\times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\Big) + \mathbf{P}_1\mathcal{M}_1\Big(\mathcal{E}\times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\Big). \qquad (6)\end{aligned}$$

By the fact $\mathbf{P}_1\mathbf{\Gamma}_1 = \mathbf{0}$, we obtain

$$\mathbf{P}_1\mathcal{M}_1\Big(\mathcal{S}\times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\Big) = \mathbf{P}_1\mathbf{G}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{A}_2^\top\widetilde{\mathbf{G}}_2^{(t-1)})\otimes(\mathbf{A}_3^\top\widetilde{\mathbf{G}}_3^{(t-1)})\big).$$

Observe that $\mathbf{P}_1\mathbf{G}_1 = \mathbf{\Phi}(\mathbf{X}_1)\mathbf{B}_1 + \mathbf{P}_1\mathbf{R}_1(\mathbf{X}_1)$. By Condition (1), we get

$$\sigma_{\min}\big(\mathbf{P}_1\mathbf{G}_1/\sqrt{I_1}\big) \geq \sigma_{\min}\big(\mathbf{\Phi}_1(\mathbf{X}_1)\mathbf{B}_1/\sqrt{I_1}\big) - O(\sqrt{R_1}\cdot J_1^{-\tau/2}) \geq 1 - O(\sqrt{R_1}\cdot J_1^{-\tau/2}).$$

Recall that the column space of $\widetilde{\mathbf{G}}_m^{(t-1)}$ is a subspace of $\mathbf{\Phi}_m(\mathbf{X}_m)$ for all $m = 1, 2, 3$, implying that $\mathbf{A}_m^\top\widetilde{\mathbf{G}}_m^{(t-1)} = \mathbf{G}_m^\top\widetilde{\mathbf{G}}_m^{(t-1)}$ and as a result

$$\sigma_{\min}\big(\mathbf{A}_m^\top\widetilde{\mathbf{G}}_m^{(t-1)}\big) = \sigma_{\min}\big(\mathbf{G}_m^\top\widetilde{\mathbf{G}}_m^{(t-1)}\big) \geq I_m\sqrt{1 - \|\widetilde{\mathbf{G}}_m^{(t-1)}\widetilde{\mathbf{G}}_m^{(t-1)\top} - \mathbf{G}_m\mathbf{G}_m^\top\|/I_m} \geq \sqrt{2}I_m/2$$

where the last inequality is due to the fact $\|\widetilde{\mathbf{G}}_m^{(t-1)}\widetilde{\mathbf{G}}_m^{(t-1)\top} - \mathbf{G}_m\mathbf{G}_m^\top\|/I_m \leq 1/2$ which holds as long as the conditions of Lemma **??** hold and initializations are warm in that $\|\widetilde{\mathbf{G}}_m^{(0)}\widetilde{\mathbf{G}}_m^{(0)\top} - \mathbf{G}_m\mathbf{G}_m^\top\|/I_m \leq 1/2$ for all $m \in [M]$. Therefore, we conclude that

$$\sigma_{\min}\Big(\mathbf{P}_1\mathcal{M}_1\big(\mathcal{S}\times_2 \widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\big)\Big) \geq \frac{\sqrt{I_1}I_2I_3}{3}\cdot\sigma_{R_1}\big(\mathcal{M}_1(\mathcal{F})\big) \geq \frac{\lambda_{\min}\sqrt{I_1}I_2I_3}{3}.$$

We now bound the operator norm of $\mathbf{P}_1\mathcal{M}_1\big(\mathcal{E}\times_2\widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3\widetilde{\mathbf{G}}_3^{(t-1)\top}\big)$. We write

$$\begin{aligned}\mathbf{P}_1\mathcal{M}_1\big(\mathcal{E}\times_2\widetilde{\mathbf{G}}_2^{(t-1)\top}\times_3\widetilde{\mathbf{G}}_3^{(t-1)\top}\big)&=\mathbf{P}_1\mathcal{M}_1(\mathcal{E}\times_2(\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3(\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})^\top)\\&+\mathbf{P}_1\mathcal{M}_1\big(\mathcal{E}\times_2(\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3(\widetilde{\mathbf{G}}_3^{(t-1)}-\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})^\top\big)\\&+\mathbf{P}_1\mathcal{M}_1\big(\mathcal{E}\times_2(\widetilde{\mathbf{G}}_2^{(t-1)}-\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3\widetilde{\mathbf{G}}_3^{(t-1)\top}\big)\end{aligned}\tag{7}$$

where $\widetilde{\mathbf{O}}_2^{(t-1)}=\arg\min_{\mathbf{O}\in\mathbb{O}^{R_2\times R_2}}\|\widetilde{\mathbf{G}}_2^{(t-1)}-\mathbf{G}_2\mathbf{O}\|_{\mathrm{F}}$ and $\widetilde{\mathbf{O}}_3^{(t-1)}=\arg\min_{\mathbf{O}\in\mathbb{O}^{R_3\times R_3}}\|\widetilde{\mathbf{G}}_3^{(t-1)}-\mathbf{G}_3\mathbf{O}\|_{\mathrm{F}}$. Clearly,

$$\begin{aligned}\big\|\mathbf{P}_1\mathcal{M}_1(\mathcal{E}\times_2(\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3(\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})^\top)\big\|&=\|\mathbf{P}_1\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{G}_3)\|\\&=\|\mathbf{\Phi}_1(\mathbf{\Phi}_1^\top\mathbf{\Phi}_1)^{-1}\mathbf{\Phi}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{G}_3)\|\end{aligned}$$

where we abuse the notation and write $\mathbf{\Phi}_1=\mathbf{\Phi}_1(\mathbf{X}_1)$. Similarly, as the proof of Lemma **??**, denote $\mathbf{U}_1$ the eigenvectors of $\mathbf{P}_1$ so that $\mathbf{U}_1^\top\mathbf{U}_1=\mathbf{I}_{J_1}$. Then,

$$\big\|\mathbf{P}_1\mathcal{M}_1(\mathcal{E}\times_2(\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3(\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})^\top)\big\|=\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{G}_3)\|$$

where the matrix $\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{G}_3)$ has size $J_1\times(R_2R_3)$.

Lemma 3. *Suppose that Assumption* **??** *holds. There exist absolute constants $C_5,C_6>0$ so that*

$$\mathbb{P}\left(\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{G}_3)\|/\sqrt{I_2I_3}\geq C_5\sqrt{J_1\log I_1}+C_6\sqrt{R_2R_3}\log^2 I_1\right)\leq 2I_1^{-2}.$$

By Lemma 3, we get that with probability at least $1-2I_1^{-2}$ that

$$\big\|\mathbf{P}_1\mathcal{M}_1(\mathcal{E}\times_2(\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3(\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})^\top)\big\|/\sqrt{I_2I_3}=O\left(\sqrt{J_1\vee(R_2R_3)}\log^2 I_1\right).\tag{8}$$

We now bound the second and third terms on RHS of (7). Write

$$\begin{aligned}&\big\|\mathbf{P}_1\mathcal{M}_1\big(\mathcal{E}\times_2(\mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top\times_3(\widetilde{\mathbf{G}}_3^{(t-1)}-\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})^\top\big)\big\|\\&=\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2\otimes(\widetilde{\mathbf{G}}_3^{(t-1)}-\mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})\big)\big\|.\end{aligned}$$

Recall $\mathbf{G}_3=\mathbf{\Phi}_3\mathbf{B}_3+\mathbf{R}_3$ where we again abused the notation and dropped their dependences

on $\mathbf{X}_3$. Therefore,

$$\Big|\|\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)}\|_{\mathrm{F}} - \|\widetilde{\mathbf{G}}_3^{t-1} - \mathbf{\Phi}_3\mathbf{B}_3\widetilde{\mathbf{O}}_3^{(t-1)}\|_{\mathrm{F}}\Big|/\sqrt{I_3} = O(\sqrt{R_3}\cdot J_3^{-\tau/2}). \tag{9}$$

We obtain

$$\begin{aligned}&\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2\otimes(\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})\big)\big\| \\ &\qquad\le\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2\otimes(\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{\Phi}_3\mathbf{B}_3\widetilde{\mathbf{O}}_3^{(t-1)})\big)\big\| + \big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2\otimes\mathbf{R}_3\big)\big\|.\end{aligned}$$

Note that the column space of $\widetilde{\mathbf{G}}_3^{(t-1)}$ belongs to the column space of $\mathbf{\Phi}_3$. Denote $\mathbf{U}_3$ the left singular vectors of $\mathbf{\Phi}_3\mathbf{B}_3$. Then,

$$\begin{aligned}&\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2\otimes(\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{\Phi}_3\mathbf{B}_3\widetilde{\mathbf{O}}_3^{(t-1)})\big)\big\| \\ &\qquad\le\|\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{\Phi}_3\mathbf{B}_3\widetilde{\mathbf{O}}_3^{(t-1)}\|_{\mathrm{F}}\cdot\sup_{\mathbf{A}\in\mathbb{R}^{J_3\times R_3},\|\mathbf{A}\|_{\mathrm{F}}\le 1}\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{U}_3\mathbf{A})\| \\ &\qquad=O\big(\|\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{\Phi}_3\mathbf{B}_3\widetilde{\mathbf{O}}_3^{(t-1)}\|_{\mathrm{F}}\sqrt{I_2}\cdot\sqrt{J_1 + J_3R_3 + R_2R_3}\log^{3/2} I_1\big),\end{aligned} \tag{10}$$

where the last inequality holds with probability at least $1 - 4I_1^{-2}$ and is due to Lemma 4.

Lemma 4. *Suppose that $\mathcal{E}$ has i.i.d entries satisfying Assumption* **??**. *Define $\mathfrak{B}(d_1, d_2) := \{\mathbf{A}\in\mathbb{R}^{d_1\times d_2}, \|\mathbf{A}\|_{\mathrm{F}}\le 1\}$. There exist absolute constants $C_1 > 0$ such that*

$$\mathbb{P}\Big(\sup_{\mathbf{A}\in\mathfrak{B}(J_3,R_3)}\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\frac{\mathbf{G}_2}{\sqrt{I_2}}\otimes\mathbf{U}_3\mathbf{A})\big\| \ge C_1\sqrt{J_1 + J_3R_3 + R_2R_3}\log^{3/2} I_1\Big) \le 4I_1^{-2} \tag{11}$$

and

$$\mathbb{P}\Bigg(\sup_{\substack{\mathbf{A}\in\mathfrak{B}(J_2,R_2)\\ \mathbf{B}\in\mathfrak{B}(J_3,R_3)}}\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\mathbf{A}\otimes\mathbf{U}_3\mathbf{B})\big\| \ge C_2\sqrt{J_1 + J_2R_2 + J_3R_3 + R_2R_3}\log^{3/2} I_1\Bigg) \le 4I_1^{-2}. \tag{12}$$

Recall that $\mathbf{U}_1, \mathbf{G}_2, \mathbf{R}_3$ are deterministic matrices. Following the same treatment as in the proof of Lemma 1, we get with probability at least $1 - 2I_1^{-2}$,

$$\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2\otimes\mathbf{R}_3\big)\big\| \le (R_3I_2I_3)^{1/2}J_3^{-\tau/2}\cdot\big(C_3\sqrt{J_1\log I_1} + C_4\sqrt{R_2R_3}\log^2 I_1\big) \tag{13}$$

for some absolute constants $C_3, C_4 > 0$.

Putting together (10) and (13), we get with probability at least $1 - 6I_1^{-2}$ that

$$\begin{aligned}&\big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})\big(\mathbf{G}_2 \otimes (\widetilde{\mathbf{G}}_3^{(t-1)} - \mathbf{G}_3\widetilde{\mathbf{O}}_3^{(t-1)})\big)\big\| \\ &\qquad \le C_5(\mathrm{Err}_{t-1} + \sqrt{R_3}\cdot J_3^{-\tau/2})\sqrt{I_2I_3}\cdot\sqrt{J_1 + J_3R_3 + R_2R_3}\log^{3/2} I_1. \end{aligned} \tag{14}$$

Similarly, we can show that with probability at least $1 - 8I_1^{-2}$,

$$\begin{aligned}&\|\mathbf{P}_1\mathcal{M}_1\big(\mathcal{E}\times_2(\widetilde{\mathbf{G}}_2^{(t-1)} - \mathbf{G}_2\widetilde{\mathbf{O}}_2^{(t-1)})^\top \times_3 \widetilde{\mathbf{G}}_3^{(t-1)\top}\big)\| \\ &\qquad \le C_6(\mathrm{Err}_{t-1} + \sqrt{R_2}\cdot J_2^{-\tau/2})\sqrt{I_2I_3}\cdot\sqrt{J_1 + J_3R_3 + R_2R_3}\log^{3/2} I_1, \end{aligned} \tag{15}$$

where we used the fact $\|\widetilde{\mathbf{G}}_3^{(t-1)}\|_{\mathrm{F}} \le \sqrt{R_3I_3}$ and the fact that the column space of $\widetilde{\mathbf{G}}_3^{(t-1)}$ is a subspace of the column space of $\mathbf{U}_3$.

Therefore, by (8), (14) and (15), we conclude that with probability at least $1 - 16I_1^{-2}$ that

$$\begin{aligned}\big\|\mathbf{P}_1\mathcal{M}_1(\mathcal{E})(\widetilde{\mathbf{G}}_2^{(t-1)} \otimes \widetilde{\mathbf{G}}_3^{(t-1)})\big\| &\le C_3\sqrt{I_2I_3}\cdot\sqrt{J_1 + R_2R_3}\log^2 I_1 \\ &+ C_6(\mathrm{Err}_{t-1} + \sqrt{R_2}J_2^{\tau/2} + \sqrt{R_3}J_3^{-\tau/2})\sqrt{I_2I_3}\cdot\sqrt{J_1 + J_3R_3 + R_2R_3}\log^{3/2} I_1.\end{aligned}$$

Now, we continue from (6). Recall that we denote $\mathring{\mathbf{G}}_1/\sqrt{I_1}$ the top-R_1 left singular vectors of $\mathbf{P}_1\mathbf{G}_1$. As shown in the proof of initialization, we have $\|\mathring{\mathbf{G}}_1\mathring{\mathbf{G}}_1^\top - \mathbf{G}_1\mathbf{G}_1^\top\|_{\mathrm{F}}/I_1 \le 2\sqrt{R_1}J_1^{-\tau/2}$. Applying Daivs-Kahan theorem to (6), we get with probability at least $1 - 16I_1^{-2}$ that

$$\begin{aligned}\|\widetilde{\mathbf{G}}_1^{(t)}\widetilde{\mathbf{G}}_1^{(t)\top} - \mathring{\mathbf{G}}_1\mathring{\mathbf{G}}_1^\top\|_{\mathrm{F}}/I_1 &\le C_4\frac{\sqrt{I_2I_3}\cdot\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1}I_2I_3} \\ &+ C_5\frac{(\mathrm{Err}_{t-1} + \sqrt{R_2}J_2^{-\tau/2} + \sqrt{R_3}J_3^{-\tau/2})\sqrt{I_2I_3}\cdot\sqrt{J_1R_1 + J_3R_1R_3 + R_1R_2R_3}\log^{3/2} I_1}{\lambda_{\min}\sqrt{I_1}I_2I_3}\end{aligned}$$

for some absolute constants $C_4, C_5 > 0$.

Therefore, as long as $\lambda_{\min}\sqrt{I_1I_2I_3} \ge C_5'\sqrt{J_1R_1 + J_3R_1R_3 + R_1R_2R_3}\log^{3/2} I_1$ for a large enough absolute constant $C_5' > 0$, we get with probability at least $1 - 16I_1^{-2}$ that

$$\|\widetilde{\mathbf{G}}_1^{(t)}\widetilde{\mathbf{G}}_1^{(t)\top} - \mathbf{G}_1\mathbf{G}_1^\top\|_{\mathrm{F}}/I_1$$

$$\leq \frac{\mathrm{Err}_{t-1}}{2} + \frac{2\sqrt{R_1}J_1^{-\tau/2} + \sqrt{R_2}J_2^{-\tau/2} + \sqrt{R_3}J_3^{-\tau/2}}{2} + C_4' \frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}}.$$

In the same fashion, we can prove similar bounds of $\|\widetilde{\mathbf{G}}_m^{(t)}\widetilde{\mathbf{G}}_m^{(t)\top} - \mathbf{G}_m\mathbf{G}_m^\top\|_{\mathrm{F}}$ for all $m = 1, 2, 3$. Therefore, with probability at least $1 - 48I_1^{-2}$,

$$\mathrm{Err}_t \leq \frac{\mathrm{Err}_{t-1}}{2} + (\sqrt{R_1}J_1^{-\tau/2} + \sqrt{R_2}J_2^{-\tau/2} + \sqrt{R_3}J_3^{-\tau/2}) + C_4' \frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}}, \tag{16}$$

which proves the first claim of Lemma **??**. The same properties can be proved for all iterations and all hold on the same event.

By the above contraction inequality in (16), after

$$t_{\max} = O(\log(\lambda_{\min}\sqrt{I_1I_2I_3/J_1}) + \tau \cdot \log(J_1) + 1)$$

iterations, we obtain

$$\mathrm{Err}_{t_{\max}} = C_4'' \frac{\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}} + \sqrt{R_1}J_1^{-\tau/2} + \sqrt{R_2}J_2^{-\tau/2} + \sqrt{R_3}J_3^{-\tau/2}$$

which holds with probability at least $1 - 48I_1^{-2}$. The proof is concluded by noting that $J_1 \asymp J_2 \asymp J_3$ and $J_1 \geq J_2 \geq J_3$. □

In order to prove Theorem **??**, we begin with proving the following result.

Lemma 5. *(Factor tensor) Suppose that conditions of Lemma* **??** *hold. Then, with probability at least* $1 - 49I_1^{-2}$ *that*

$$\begin{aligned}&\|\widetilde{\mathcal{F}} - \mathcal{F} \times_1 \widetilde{\mathbf{O}}_1^\top \times_2 \widetilde{\mathbf{O}}_2^\top \times_3 \widetilde{\mathbf{O}}_3^\top\|_{\mathrm{F}} \\ &\quad\leq C_1\Big(\frac{\kappa_0 \cdot \sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\sqrt{I_1I_2I_3}}\Big) + 2\kappa_0\lambda_{\min}\sqrt{R_1}J_1^{-\tau/2}.\end{aligned}$$

where $\widetilde{\mathbf{O}}_m$ *is an orthogonal matrix which realizes the minimium* $\min_{\mathbf{O}} \|\widetilde{\mathbf{G}}_m - \mathbf{G}_m\mathbf{O}\|_{\mathrm{F}}$ *and* $C_1 > 0$ *is an absolute constant.*

Proof of Lemma 5. Recall that $\widetilde{\mathcal{F}} = (I_1I_2I_3)^{-1} \cdot \mathcal{Y} \times_1 \widetilde{\mathbf{G}}_1^\top \times_2 \widetilde{\mathbf{G}}_2^\top \times_3 \widetilde{\mathbf{G}}_3^\top$ and so that

$$\widetilde{\mathcal{F}} = (I_1I_2I_3)^{-1} \cdot \mathcal{F} \times_1 (\widetilde{\mathbf{G}}_1^\top \mathbf{G}_1) \times_2 (\widetilde{\mathbf{G}}_2^\top \mathbf{G}_2) \times_3 (\widetilde{\mathbf{G}}_3^\top \mathbf{G}_3) + (I_1I_2I_3)^{-1} \cdot \mathcal{E} \times_1 \widetilde{\mathbf{G}}_1^\top \times_2 \widetilde{\mathbf{G}}_2^\top \times_3 \widetilde{\mathbf{G}}_3^\top$$

where we used the fact $\widetilde{\mathbf{G}}_m^\top \mathbf{\Gamma}_m = \mathbf{0}$ since the column space of $\widetilde{\mathbf{G}}_m$ is a subspace of the column space of $\mathbf{\Phi}_m(\mathbf{X}_m)$. Recall that

$$\mathbf{G}_m^\top \big(\widetilde{\mathbf{G}}_m \widetilde{\mathbf{G}}_m^\top - \mathbf{G}_m \mathbf{G}_m^\top\big) \mathbf{G}_m / I_m^2 = \mathbf{G}_m^\top \widetilde{\mathbf{G}}_m (\mathbf{G}_m^\top \widetilde{\mathbf{G}}_m)^\top / I_m^2 - \mathbf{I}_{R_m}$$

where $\mathbf{G}_m^\top \widetilde{\mathbf{G}}_m$ is an $R_m \times R_m$ matrix. Then, by Lemma **??**, with probability at least $1 - 48 I_1^{-2}$ that

$$\big\| \mathbf{G}_m^\top \widetilde{\mathbf{G}}_m (\mathbf{G}_m^\top \widetilde{\mathbf{G}}_m)^\top / I_m^2 - \mathbf{I}_{R_m} \big\|_{\mathrm{F}} \le C_5 \frac{\sqrt{J_1 R_1 + R_1 R_2 R_3} \log^2 I_1}{\lambda_{\min} \sqrt{I_1 I_2 I_3}} + 2\sqrt{R_1} J_1^{-\tau/2}.$$

It implies that for all $m = 1, 2, 3$, there exists an orthonormal matrix $\widetilde{\mathbf{O}}_m \in \mathbb{O}^{R_m \times R_m}$ so that

$$\| \widetilde{\mathbf{G}}_m^\top \mathbf{G}_m / I_m - \widetilde{\mathbf{O}}_m^\top \|_{\mathrm{F}} = C_5 \frac{\sqrt{J_1 R_1 + R_1 R_2 R_3} \log^2 I_1}{\lambda_{\min} \sqrt{I_1 I_2 I_3}} + 2\sqrt{R_1} J_1^{-\tau/2},$$

which holds with the same probability. Therefore,

$$\begin{aligned} \widetilde{\mathcal{F}} - \mathcal{F} \times_1 \widetilde{\mathbf{O}}_1^\top \times_2 \widetilde{\mathbf{O}}_2^\top \times_3 \widetilde{\mathbf{O}}_3^\top &= (I_1I_2I_3)^{-1} \cdot \mathcal{E} \times_1 \widetilde{\mathbf{G}}_1^\top \times_2 \widetilde{\mathbf{G}}_2^\top \times_3 \widetilde{\mathbf{G}}_3^\top \\ &\quad + \mathcal{F}\big(\times_1 (\widetilde{\mathbf{G}}_1^\top \mathbf{G}_1 / I_1) \times_2 (\widetilde{\mathbf{G}}_2^\top \mathbf{G}_2 / I_2) \times_3 (\widetilde{\mathbf{G}}_3^\top \mathbf{G}_3 / I_3) - \times_1 \widetilde{\mathbf{O}}_1^\top \times_2 \widetilde{\mathbf{O}}_2^\top \times_3 \widetilde{\mathbf{O}}_3^\top \big). \end{aligned} \tag{17}$$

Observe that

$$\begin{aligned} &\big\| \mathcal{F} \times_1 (\widetilde{\mathbf{G}}_1^\top \mathbf{G}_1 / I_1 - \widetilde{\mathbf{O}}_1^\top) \times_2 (\widetilde{\mathbf{G}}_2^\top \mathbf{G}_2 / I_2) \times_3 (\widetilde{\mathbf{G}}_3^\top \mathbf{G}_3 / I_3) \big\|_{\mathrm{F}} \\ &\quad \le \big\| (\widetilde{\mathbf{G}}_1^\top \mathbf{G}_1 / I_1 - \widetilde{\mathbf{O}}_1^\top) \mathcal{M}_1(\mathcal{F}) \big((\widetilde{\mathbf{G}}_2^\top \mathbf{G}_2 / I_2) \otimes (\widetilde{\mathbf{G}}_3^\top \mathbf{G}_3 / I_3) \big) \big\|_{\mathrm{F}} \\ &\quad \le \| \mathcal{M}_1(\mathcal{F}) \| \cdot \| \widetilde{\mathbf{G}}_1^\top \mathbf{G}_1 / I_1 - \widetilde{\mathbf{O}}_1^\top \|_{\mathrm{F}} \\ &\quad \le C_5 \frac{\kappa_0 \cdot \sqrt{J_1 R_1 + R_1 R_2 R_3} \log^2 I_1}{\sqrt{I_1 I_2 I_3}} + 2\kappa_0 \lambda_{\min} \sqrt{R_1} J_1^{-\tau/2}. \end{aligned}$$

As a result, we can show with probability at least $1-48I_1^{-2}$ that

$$\begin{aligned}&\big\|\mathcal{F}\big(\times_1(\widetilde{\mathbf{G}}_1^\top\mathbf{G}_1/I_1)\times_2(\widetilde{\mathbf{G}}_2^\top\mathbf{G}_2/I_2)\times_3(\widetilde{\mathbf{G}}_3^\top\mathbf{G}_3/I_3)-\times_1\widetilde{\mathbf{O}}_1^\top\times_2\widetilde{\mathbf{O}}_2^\top\times_3\widetilde{\mathbf{O}}_3^\top\big)\big\|\\&\le C_5'\frac{\kappa_0\cdot\sqrt{J_1R_1+R_1R_2R_3}\log^2 I_1}{\sqrt{I_1I_2I_3}}+6\kappa_0\lambda_{\min}\sqrt{R_1}J_1^{-\tau/2}.\end{aligned}\tag{18}$$

Observe that the rank of $\mathcal{M}_1\big(\mathcal{E}\times_1\widetilde{\mathbf{G}}_1^\top\times_2\widetilde{\mathbf{G}}_2^\top\times_3\widetilde{\mathbf{G}}_3^\top\big)$ is bounded by R_1. Similarly,

$$\begin{aligned}\big\|\mathcal{E}\times_1\widetilde{\mathbf{G}}_1^\top\times_2\widetilde{\mathbf{G}}_2^\top\times_3\widetilde{\mathbf{G}}_3^\top\big\|_{\mathrm{F}}&=\big\|\mathcal{M}_1\big(\mathcal{E}\times_1\widetilde{\mathbf{G}}_1^\top\times_2\widetilde{\mathbf{G}}_2^\top\times_3\widetilde{\mathbf{G}}_3^\top\big)\big\|_{\mathrm{F}}\\&\le\sqrt{R_1}\cdot\big\|\mathcal{M}_1\big(\mathcal{E}\times_1\widetilde{\mathbf{G}}_1^\top\times_2\widetilde{\mathbf{G}}_2^\top\times_3\widetilde{\mathbf{G}}_3^\top\big)\big\|.\end{aligned}$$

Lemma 6. *Suppose that Assumption ?? holds and assume $J_1\asymp J_2\asymp J_3$ and $J_1\ge R_1\ge R_2\ge R_3$. There exist absolute constants $C_7>0$ so that,*

$$\mathbb{P}\Big(\big\|\mathcal{M}_1\big(\mathcal{E}\times_1\widetilde{\mathbf{G}}_1^\top\times_2\widetilde{\mathbf{G}}_2^\top\times_3\widetilde{\mathbf{G}}_3^\top\big)\big\|/\sqrt{I_1I_2I_3}\ge C_7\sqrt{J_1R_1}\log^{3/2}I_1\Big)\le I_1^{-2}.$$

By Lemma 6, with probability at least $1-I_1^{-2}$,

$$\big\|\mathcal{M}_1\big(\mathcal{E}\times_1\widetilde{\mathbf{G}}_1^\top\times_2\widetilde{\mathbf{G}}_2^\top\times_3\widetilde{\mathbf{G}}_3^\top\big)\big\|/\sqrt{I_1I_2I_3}\le C_1'\sqrt{J_1R_1}\log^{3/2}I_1.\tag{19}$$

for some absolute constant $C_1'>0$.

Putting together (17), (18) and (19), we conclude that with probability at least $1-49I_1^{-2}$ that

$$\begin{aligned}&\|\widetilde{\mathcal{F}}-\mathcal{F}\times_1\widetilde{\mathbf{O}}_1^\top\times_2\widetilde{\mathbf{O}}_2^\top\times_3\widetilde{\mathbf{O}}_3^\top\|_{\mathrm{F}}\\&\qquad\le C_8\frac{\kappa_0\cdot\sqrt{J_1R_1+R_1R_2R_3}\log^2 I_1}{\sqrt{I_1I_2I_3}}+6\kappa_0\lambda_{\min}\sqrt{R_1}J_1^{-\tau/2},\end{aligned}$$

which proves Lemma 5. □

Proof of Theorem ??. Let $\widehat{\mathbf{O}}_m$ denote the left singular vectors of $\mathcal{M}_m(\widetilde{\mathcal{F}})$ for all $m\in[M]$, and $\widetilde{\mathbf{D}}_m$ denote the singular values of $\mathcal{M}_m(\widetilde{\mathcal{F}})$. Similarly, denote $\mathbf{D}_m$ the singular values of $\mathcal{M}_m(\mathcal{F})$. Lemma 5 implies, with probability at least $1-49I_1^{-2}$, that

$$\|\mathbf{D}_m-\widetilde{\mathbf{D}}_m\|\le C_8\frac{\kappa_0\cdot\sqrt{J_1R_1+R_1R_2R_3}\log^2 I_1}{\sqrt{I_1I_2I_3}}+6\kappa_0\lambda_{\min}\sqrt{R_1}J_1^{-\tau/2}$$

and

$$\begin{aligned}&\|\widetilde{\mathbf{O}}_1\mathcal{M}_1(\widetilde{\mathcal{F}})\mathcal{M}_1(\widetilde{\mathcal{F}})^\top\widetilde{\mathbf{O}}_1^\top-\mathcal{M}_1(\mathcal{F})\mathcal{M}_1(\mathcal{F})^\top\|\\&\quad=\|\mathcal{M}_1(\widetilde{\mathcal{F}})\mathcal{M}_1(\widetilde{\mathcal{F}})^\top-\widetilde{\mathbf{O}}_1^\top\mathcal{M}_1(\mathcal{F})\mathcal{M}_1(\mathcal{F})^\top\widetilde{\mathbf{O}}_1\|.\end{aligned}$$

Denote $\widetilde{\mathbf{H}}_1=\widetilde{\mathbf{O}}_1\widehat{\mathbf{O}}_1$ so that $\widetilde{\mathbf{O}}_1\mathcal{M}_1(\widetilde{\mathcal{F}})\mathcal{M}_1(\widetilde{\mathcal{F}})^\top\widetilde{\mathbf{O}}_1^\top-\mathcal{M}_1(\mathcal{F})\mathcal{M}_1^\top(\mathcal{F})=\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1^\top-\mathbf{D}_1^2$ where we used Assumption **??**. Denote $\varepsilon_\alpha=C_8\kappa_0^2\lambda_{\min}\sqrt{J_1R_1+R_1R_2R_3}\log^2(I_1)/\sqrt{I_1I_2I_3}+6\kappa_0^2\lambda_{\min}^2\sqrt{R_1}J_1^{-\tau/2}$. Then, by Lemma 5, we obtain with probability at least $1-49I_1^{-2}$ that

$$\|\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1^\top-\mathbf{D}_1^2\|_{\mathrm{F}}\le\varepsilon_\alpha. \tag{20}$$

Note that for each $j=1,\cdots,R_1$, we obtain $\sigma_j(\mathbf{D}_1^2)-\sigma_{j+1}(\mathbf{D}_1^2)\ge\lambda_{\min}\cdot\mathrm{Egap}(\mathcal{F})$. Under the conditions of Theorem **??**, it follows with probability at least $1-49I_1^{-2}$ that

$$\min_{1\le j\le R_1}\sigma_j(\mathbf{D}_1^2)-\sigma_{j+1}(\mathbf{D}_1^2)\ge C_1\kappa_0^2\sqrt{R_1}\cdot\varepsilon_\alpha$$

for a large enough constant $C_1>1$ implying that the order of eigenvalues of $\mathbf{D}_1^2$ will be maintained in view of (20). By applying the Davis-Kahan theorem to each isolated eigenvector of $\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1$, we can conclude that $\|\widetilde{\mathbf{h}}_j\widetilde{\mathbf{h}}_j^\top-\mathbf{e}_j\mathbf{e}_j^\top\|\le1/(2\kappa_0^2\sqrt{R_1})$ which holds for all $j=1,\cdots,R_1$ where $\widetilde{\mathbf{h}}_j$ denotes the j-th column of $\widetilde{\mathbf{H}}_1$ and $\mathbf{e}_j$ denotes the j-th canonical basis vector. Indeed, it holds as long as the $\mathrm{Egap}(\mathcal{F})$ is large enough as stated in the conditions of Theorem **??**. It implies that there exists a $\widetilde{s}_j\in\{\pm1\}$ so that $\|\widetilde{\mathbf{h}}_j\widetilde{s}_j-\mathbf{e}_j\|\le1/\sqrt{2\kappa_0^4R_1}$ for each j. Denote $\widetilde{\mathbf{S}}_1=\mathrm{diag}(\widetilde{s}_1,\cdots,\widetilde{s}_{R_1})$ so that

$$\|\widetilde{\mathbf{H}}_1\widetilde{\mathbf{S}}_1-\mathbf{I}_{R_1}\|_{\mathrm{F}}\le\Big(\sum_{j=1}^{R_1}\|\widetilde{\mathbf{h}}_j\widetilde{s}_j-\mathbf{e}_j\|^2\Big)^{1/2}\le1/(\sqrt{2}\kappa_0^2).$$

Note that, on the same event, $\|\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1^\top-\widetilde{\mathbf{D}}_1^2\|_{\mathrm{F}}\le\varepsilon_\alpha+\|\widetilde{\mathbf{D}}_1^2-\mathbf{D}_1^2\|_{\mathrm{F}}\le2\varepsilon_\alpha$ where the last bound is due to Lemma 5. Since $\widetilde{\mathbf{D}}_1$ is a diagonal matrix, $\|\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1^\top-\widetilde{\mathbf{S}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\le2\varepsilon_\alpha$. Write

$$\|\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1^\top-\widetilde{\mathbf{S}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{S}}_1^\top\|_{\mathrm{F}}\ge\|(\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1)\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{S}}_1^\top+\widetilde{\mathbf{S}}_1\widetilde{\mathbf{D}}_1^2(\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1)^\top\|_{\mathrm{F}}$$

$$
\begin{aligned}
-\|(\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1)\widetilde{\mathbf{D}}_1^2(\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1)^\top\|_{\mathrm{F}} &\geq 2\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\sigma_{\min}(\widetilde{\mathbf{D}}_1^2)-O\big(\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}^2\sigma_{\max}(\widetilde{\mathbf{D}}_1^2)\big)\\
&\geq 2\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\sigma_{\min}(\widetilde{\mathbf{D}}_1^2)-O\big(\kappa_0^2\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}^2\sigma_{\min}(\widetilde{\mathbf{D}}_1^2)\big)
\end{aligned}
$$

where the last inequality holds with probability at least $1-49I_1^{-2}$ as long as $\|\mathbf{D}_1-\widetilde{\mathbf{D}}_1\|\leq\lambda_{\min}/4$ which is guaranteed by the lower bound on $\lambda_{\min}$. It implies that

$$
\|\widetilde{\mathbf{H}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{H}}_1^\top-\widetilde{\mathbf{S}}_1\widetilde{\mathbf{D}}_1^2\widetilde{\mathbf{S}}_1^\top\|_{\mathrm{F}}\geq(2-\sqrt{2})\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\sigma_{\min}(\widetilde{\mathbf{D}}_1^2)\geq\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\lambda_{\min}^2/5.
$$

Therefore, we conclude with probability at least $1-49I_1^{-2}$ that

$$
\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\leq 10\varepsilon_\alpha/\lambda_{\min}^2\leq C_7\frac{\kappa_0^2\sqrt{JR_1+R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}}+C_8\kappa_0^2\sqrt{R_1}J_1^{-\tau/2}.
$$

As a result, $\widehat{\mathbf{G}}_1=\widetilde{\mathbf{G}}_1\widehat{\mathbf{O}}_1$ and then with probability at least $1-49I_1^{-2}$,

$$
\begin{aligned}
\|\widehat{\mathbf{G}}_1-\mathbf{G}_1\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}/\sqrt{I_1}\leq&\|\widehat{\mathbf{G}}_1-\mathbf{G}_1\widetilde{\mathbf{H}}_1\|_{\mathrm{F}}/\sqrt{I_1}+\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\\
=&\|\widehat{\mathbf{G}}_1-\mathbf{G}_1\widetilde{\mathbf{O}}_1\widehat{\mathbf{O}}_1\|_{\mathrm{F}}/\sqrt{I_1}+\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\\
=&\|\widetilde{\mathbf{G}}_1-\mathbf{G}_1\widetilde{\mathbf{O}}_1\|_{\mathrm{F}}/\sqrt{I_1}+\|\widetilde{\mathbf{H}}_1-\widetilde{\mathbf{S}}_1\|_{\mathrm{F}}\\
\leq&C_7\frac{\kappa_0^2\sqrt{JR_1+R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}}+C_8\kappa_0^2\sqrt{R_1}J_1^{-\tau/2}
\end{aligned}
$$

where the last inequality is due to that $\widetilde{\mathbf{O}}_1$ realizes the minimum of $\min_{\mathbf{O}}\|\widetilde{\mathbf{G}}_1-\mathbf{G}_1\mathbf{O}\|_{\mathrm{F}}$. Clearly, the bounds can be proved identically for all $\|\widehat{\mathbf{G}}_m-\mathbf{G}_m\mathbf{S}_m\|_{\mathrm{F}}/\sqrt{I_m}$.

At last, recall that $\widehat{\mathcal{F}}=\widetilde{\mathcal{F}}\times_1\widehat{\mathbf{O}}_1^\top\times_2\widehat{\mathbf{O}}_2^\top\times_3\widehat{\mathbf{O}}_3^\top$. We conclude that with probability at least $1-49I_1^{-2}$,

$$
\begin{aligned}
\|\widehat{\mathcal{F}}-\mathcal{F}\times_1\mathbf{S}_1\times_2\mathbf{S}_2\times_3\mathbf{S}_3\|_{\mathrm{F}}=&\|\widehat{\mathcal{F}}-\mathcal{F}\times_1\widetilde{\mathbf{H}}_1^\top\times_2\widetilde{\mathbf{H}}_2^\top\times_3\widetilde{\mathbf{H}}_3^\top\|_{\mathrm{F}}+O(\kappa_0\varepsilon_\alpha/\lambda_{\min})\\
=&\|\widetilde{\mathcal{F}}\times_1\widehat{\mathbf{O}}_1^\top\times_2\widehat{\mathbf{O}}_2^\top\times_3\widehat{\mathbf{O}}_3^\top-\mathcal{F}\times_1\widetilde{\mathbf{H}}_1^\top\times_2\widetilde{\mathbf{H}}_2^\top\times_3\widetilde{\mathbf{H}}_3^\top\|_{\mathrm{F}}+O(\kappa_0\varepsilon_\alpha/\lambda_{\min})\\
=&\|\widetilde{\mathcal{F}}-\mathcal{F}\times_1(\widehat{\mathbf{O}}_1\widetilde{\mathbf{H}}_1^\top)\times_2(\widehat{\mathbf{O}}_2\widetilde{\mathbf{H}}_2^\top)\times_3(\widehat{\mathbf{O}}_3\widetilde{\mathbf{H}}_3^\top)\|_{\mathrm{F}}+O(\kappa_0\varepsilon_\alpha/\lambda_{\min})\\
=&\|\widetilde{\mathcal{F}}-\mathcal{F}\times_1\widetilde{\mathbf{O}}_1^\top\times_2\widetilde{\mathbf{O}}_2^\top\times_3\widetilde{\mathbf{O}}_3^\top\|_{\mathrm{F}}+O(\kappa_0\varepsilon_\alpha/\lambda_{\min})=O((\kappa_0\varepsilon_\alpha/\lambda_{\min})\\
=&O\Big(\frac{\kappa_0^3\sqrt{JR_1+R_1R_2R_3}\log^2 I_1}{\sqrt{I_1I_2I_3}}+\kappa_0^3\lambda_{\min}\sqrt{R_1}J_1^{-\tau/2}\Big),
\end{aligned}
$$

which proves Theorem **??**. □

Proof of Theorem **??**. Without loss of generality, we prove the bound for $m = 1$. Recall by definition that

$$\widehat{\mathbf{\Gamma}}_1 = \mathbf{P}_1^{\perp}\widehat{\mathbf{A}}_1 = \mathbf{P}_1^{\perp}\mathcal{M}_1(\mathcal{Y} \times_2 \mathbf{P}_2 \times_2 \mathbf{P}_3)\mathcal{M}_1(\widehat{\mathcal{F}} \times_2 \widehat{\mathbf{G}}_2 \times_3 \widehat{\mathbf{G}}_3)^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3).$$

Since the column space of $\widehat{\mathbf{G}}_m$ is a subspace of the column space of $\mathbf{P}_m$ so that $\widehat{\mathbf{G}}_m^{\top}\mathbf{\Gamma}_m = \mathbf{0}$, we can write

$$\begin{aligned}\mathcal{M}_1(\mathcal{Y} \times_2 \mathbf{P}_2 \times_3 \mathbf{P}_3)\mathcal{M}_1(\widehat{\mathcal{F}} \times_2 \widehat{\mathbf{G}}_2 \times_3 \widehat{\mathbf{G}}_3)^{\top} &= \mathcal{M}_1(\mathcal{Y})(\mathbf{P}_2 \otimes \mathbf{P}_3)(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top} \\ &= \mathcal{M}_1(\mathcal{Y})(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top} \\ &= (\mathbf{G}_1 + \mathbf{\Gamma}_1)\mathcal{M}_1(\mathcal{F})\big((\mathbf{G}_2^{\top}\widehat{\mathbf{G}}_2) \otimes (\mathbf{G}_3^{\top}\widehat{\mathbf{G}}_3)\big)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top} + \mathcal{M}_1(\mathcal{E})(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\end{aligned}$$

and as a result

$$\begin{aligned}\widehat{\mathbf{\Gamma}}_1 =& \mathbf{P}_1^{\perp}\mathbf{\Gamma}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{G}_2^{\top}\widehat{\mathbf{G}}_2) \otimes (\mathbf{G}_3^{\top}\widehat{\mathbf{G}}_3)\big)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3) \\ &+ \mathbf{P}_1^{\perp}\mathcal{M}_1(\mathcal{E})(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3).\end{aligned}$$

Under the conditions of Lemma **??** and by Theorem **??**, we conclude with probability at least $1 - 49I_1^{-2}$ that $\sigma_{\min}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\big) \geq \lambda_{\min}/2$. Now, it suffices to bound the spectral norm $\mathbf{P}_1^{\perp}\mathcal{M}_1(\mathcal{E})(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}$. Since the column spaces of $\widehat{\mathbf{G}}_2$ and $\widehat{\mathbf{G}}_3$ are the subspaces of column spaces of $\mathbf{\Phi}_2(\mathbf{X}_2)$ and $\mathbf{\Phi}_3(\mathbf{X}_3)$, respectively, we have

$$\begin{aligned}&\big\|\mathbf{P}_1^{\perp}\mathcal{M}_1(\mathcal{E})(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3)\big\| \\ \leq&\big\|\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}\big\|/(I_2I_3) \cdot \sup_{\substack{\mathbf{A}_2 \in \mathfrak{B}(J_2,R_2), \mathbf{A}_3 \in \mathfrak{B}(J_3,R_3) \\ \mathbf{B} \in \mathfrak{B}(R_2R_3,R_1)}} \|\mathbf{P}_1^{\perp}\mathcal{M}_1(\mathcal{E})\big((\mathbf{U}_2\mathbf{A}_2) \otimes (\mathbf{U}_3\mathbf{A}_3)\big)\mathbf{B}\|\end{aligned}$$

where $\mathbf{U}_m$ are the left singular vectors of $\mathbf{\Phi}_m(\mathbf{X}_m)$ and $\mathfrak{B}(d_1, d_2) = \{\mathbf{B} \in \mathbb{R}^{d_1 \times d_2} : \|\mathbf{B}\| \leq 1\}$. The following lemma is needed whose proof is reproducible by the proof of Lemma 6.

Lemma 7. *Suppose that Assumption* **??** *holds and assume* $J_1 \asymp J_2 \asymp J_3$ *and* $J_1 \geq R_1 \geq$

$R_2 \geq R_3$. *There exist an absolute constant* $C_9 > 0$ *so that with probability at least* $1 - I_1^{-2}$,

$$\sup_{\substack{\mathbf{A}_2 \in \mathfrak{B}(J_2,R_2), \mathbf{A}_3 \in \mathfrak{B}(J_3,R_3) \\ \mathbf{B} \in \mathfrak{B}(R_2R_3,R_1)}} \|\mathbf{P}_1^{\perp} \mathcal{M}_1(\mathcal{E})\big((\mathbf{U}_2\mathbf{A}_2) \otimes (\mathbf{U}_3\mathbf{A}_3)\big)\mathbf{B}\| \leq C_9 \sqrt{I_1 + J_1R_1 + R_1R_2R_3} \log^{3/2} I_1.$$

By Lemma 7, we get with probability at least $1 - I_1^{-2}$ that

$$\begin{aligned} &\Big\|\mathbf{P}_1^{\perp} \mathcal{M}_1(\mathcal{E})(\widehat{\mathbf{G}}_2 \otimes \widehat{\mathbf{G}}_3)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3)\Big\|_{\mathrm{F}} \\ &\qquad \leq C_9' \frac{\sqrt{I_1R_1 + J_1R_1^2 + R_1^2R_2R_3} \log^{3/2} I_1}{\lambda_{\min}\sqrt{I_2I_3}} \end{aligned}$$

where we used the fact $\big\|\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}\big\| \leq C_1'\lambda_{\min}^{-1}$ by Theorem **??** and conditions of Lemma **??**.

Since $\mathbf{P}_1^{\perp}\mathbf{\Gamma}_1 = \mathbf{\Gamma}_1$, we get

$$\begin{aligned} &\mathbf{P}_1^{\perp}\mathbf{\Gamma}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{G}_2^{\top}\widehat{\mathbf{G}}_2) \otimes (\mathbf{G}_3^{\top}\widehat{\mathbf{G}}_3)\big)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3) \\ =&\mathbf{\Gamma}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{G}_2^{\top}\widehat{\mathbf{G}}_2) \otimes (\mathbf{G}_3^{\top}\widehat{\mathbf{G}}_3)\big)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1}/(I_2I_3) \\ =&\mathbf{\Gamma}_1\mathcal{M}_1(\mathcal{F})\big(\mathbf{S}_2 \otimes \mathbf{S}_3\big)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1} \\ &+ O\Big(\kappa_0\|\mathbf{\Gamma}_1\| \cdot (\|\widehat{\mathbf{G}}_2^{\top}\mathbf{G}_2/I_2 - \mathbf{S}_2\|_{\mathrm{F}} + \|\widehat{\mathbf{G}}_3^{\top}\mathbf{G}_3/I_3 - \mathbf{S}_3\|_{\mathrm{F}})\Big) \end{aligned}$$

where the last term is bounded in terms of Frobenius norm and $\mathbf{S}_2, \mathbf{S}_3$ are defined as in Theorem **??**. Meanwhile,

$$\big\|\mathbf{S}_1^{\top}\mathcal{M}_1(\widehat{\mathcal{F}}) - \mathcal{M}_1(\mathcal{F})\big(\mathbf{S}_2^{\top} \otimes \mathbf{S}_3^{\top}\big)\big\|_{\mathrm{F}} \leq \|\widehat{\mathcal{F}} - \mathcal{F} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \times_3 \mathbf{S}_3\|_{\mathrm{F}}.$$

Therefore, by Theorem **??**, with probability at least $1 - 49I_1^{-2}$ that

$$\begin{aligned} &\Big\|\mathbf{\Gamma}_1\mathcal{M}_1(\mathcal{F})\big(\mathbf{S}_2^{\top} \otimes \mathbf{S}_3^{\top}\big)\mathcal{M}_1(\widehat{\mathcal{F}})^{\top}\big(\mathcal{M}_1(\widehat{\mathcal{F}})\mathcal{M}_1^{\top}(\widehat{\mathcal{F}})\big)^{-1} - \mathbf{\Gamma}_1\mathbf{S}_1^{\top}\Big\|_{\mathrm{F}} \\ &\qquad = O\big(\lambda_{\min}^{-1}\|\mathbf{\Gamma}_1\| \cdot \|\widehat{\mathcal{F}} - \mathcal{F} \times_1 \mathbf{S}_1 \times_2 \mathbf{S}_2 \times_3 \mathbf{S}_3\|_{\mathrm{F}}\big) \\ = O\Big(\|\mathbf{\Gamma}_1\| \cdot &\frac{\kappa_0^3\sqrt{J_1R_1 + R_1R_2R_3} \log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}}\Big) + O\big(\|\mathbf{\Gamma}_1\| \cdot \kappa_0^3\sqrt{R_1}J_1^{-\tau/2}\big). \end{aligned}$$

Finally, we get with probability at least $1 - 50I_1^{-2}$ that

$$\begin{aligned}&\|\widehat{\mathbf{\Gamma}}_1 - \mathbf{\Gamma}_1\mathbf{S}_1^\top\|_{\mathrm{F}}\\ &=O\Big(\|\mathbf{\Gamma}_1\| \cdot \Big(\frac{\kappa_0^3\sqrt{J_1R_1 + R_1R_2R_3}\log^2 I_1}{\lambda_{\min}\sqrt{I_1I_2I_3}} + \kappa_0^3\sqrt{R_1}J_1^{-\tau/2}\Big)\Big) + O\Big(\frac{\sqrt{R_1I_1 + J_1R_1^2 + R_1^2R_2R_3}\log^{3/2} I_1}{\lambda_{\min}\sqrt{I_2I_3}}\Big)\end{aligned}$$

which concludes the proof of Theorem **??** in view of $J_m \leq I_m$. □

Appendix B Number of Basis Functions

Determination of the number of basis functions is an important task in non-parametric and semi-parametric estimations. It is more challenging in the STEFA model. According to our analysis in "Effect of the number of fitting basis" in the simulation section, the interactions between the true number of basis, the working number of basis, the signal-to-noise ratio, and the relative mean squared errors is not straightforward. Specifically, Table **??** shows that increasing the sieve order J does not always improve the performance and $J = 16$ does not achieve the best performance among all choices of J, even though the data is simulated with order 16.

To start a formal investigation of this challenging problem, we can first take the perspective of a regression problem:

$$\mathcal{M}_m(\mathcal{Y}) = \mathbf{\Phi}_m\mathbf{B}_m + \mathbf{E}_m,$$

where $\mathbf{\Phi}_m$ is an ensemble of basis functions, $\mathbf{B}_m$ is the coefficients and $\mathbf{E}_m$ is the residual. An potential data-driven way to determine the sieve degree is to construct an F-test based on the statistics $(\|P_{\Phi_m}\mathcal{M}_m(\mathcal{Y})\|_F^2 - \|P_{\Phi'_m}\mathcal{M}_m(\mathcal{Y})\|_F^2)/\|P_{\Phi_m}^{\perp}\mathcal{M}_m(\mathcal{Y})\|_F^2$ when comparing two choices of sieve degrees (corresponding to Φ_m and the reduced one Φ'_m). However, strictly speaking, the residual $\mathbf{E}_m$ is not of multivariate Gaussian and the coefficient matrix $\mathbf{B}_m$ is restricted to a certain low rank structure due to the other modes in tensor $\mathcal{Y}$. The proper test for this sieve determination needs further investigation and is beyond the current main streamline of this paper.

In the simulation and real data analysis sections, we choose the sieve degree J in an

ad-hoc way in the simulation. For instance the degree used in polynomial basis and B-spline basis are chosen to accommodate one's expectation on the smoothness of the function. The impact of such ad-hoc choices of sieve degree J was investigated in Table 3 in the simulation section, where an obvious bias-variance trade-off was observed. In real data analyses, we choose the J that minimizes the relative mean squared errors.

Here, we present some additional empirical results of selecting the number of basis through relative mean squared errors (ReMSE). In this simulation, we consider a three-way tensor with fixed dimensions $I_1 = I_2 = I_3 = 100$, whose signal part can be decomposed to a Tucker decomposition with rank $R = (3, 3, 3)$. We fix the signal-to-noise ratio $\alpha = 1.5$, and simulate the parametric part of loading within the manifold space of Legendre function of a two-dimensional $\mathbf{X}_m, m = 1, 2, 3$. The magnitude of $\mathbf{\Gamma}_m$ is controlled by μ as in Section 5. We consider three different magnitudes of $\mathbf{\Gamma}_m$'s and four different numbers of true basis J. For each combination of (μ, J), we simulate for 100 times and report the average number of selected basis in Table 1 for four different methods, which minimize (a) in-sample ReMSE$_{\mathcal{Y}}$; (b) in-sample ReMSE$_{\mathcal{S}}$; (c) out-of-sample ReMSE$_{\mathcal{Y}}$; and (d) out-of-sample ReMSE$_{\mathcal{S}}$, respectively. We note that the ReMSE with respect to signal ReMSE$_{\mathcal{S}}$ is only available in simulation environment and the ReMSE with respect to observed tensors ReMSE$_{\mathcal{Y}}$ is more practical in real data analysis.

Table 1 shows that selecting number of basis by minimizing in-sample ReMSE usually leads to an over-estimation of J. However, selecting number of basis by minimizing out-of-sample ReMSE usually produces more accurate results. Comparing the last two columns, we notice that, as long as we use the out-of-sample ReMSE, it does not really matter whether we use ReMSE with respect to observed noisy $\mathcal{Y}$ or the true signal $\mathcal{S}$ for all different combinations of magnitudes of $\mathbf{\Gamma}_m$ and number of true basis J. This observation provides empirical support to using out-of-sample ReMSE$_{\mathcal{Y}}$ to select J in real applications where ReMSE$_{\mathcal{S}}$ can not be calculated. Moreover, by comparing estimated $\widehat{J}$ by minimizing out-of-sample ReMSE across different true (μ, J) combinations, we observe that out-of-sample $\widehat{J}$ tends to give an under-estimation of the true number of basis for the purpose of robustness when J is large.

Truth		Average $\widehat{J}$ by minimizing In-Sample		Average $\widehat{J}$ by minimizing Out-of-Sample	
μ	J	$\text{ReMSE}_{\mathcal{Y}}$	$\text{ReMSE}_{\mathcal{S}}$	$\text{ReMSE}_{\mathcal{Y}}$	$\text{ReMSE}_{\mathcal{S}}$
0.1	2	3.97	3.31	2.00	2.00
0.1	4	8.00	6.52	4.00	4.00
0.1	8	16.00	13.70	8.02	8.02
0.1	16	32.00	28.32	11.38	11.38
0.3	2	3.97	3.76	2.00	2.00
0.3	4	7.99	7.62	4.00	4.00
0.3	8	16.00	15.60	7.81	7.81
0.3	16	32.00	31.60	8.35	8.34
0.5	2	3.99	3.96	2.00	2.00
0.5	4	8.00	7.90	4.00	4.00
0.5	8	15.99	15.93	6.79	6.78
0.5	16	32.00	31.86	6.88	6.88

Table 1: Average number of $\widehat{J}$ of selected basis by the four methods over 100 repetitions.

Appendix C Kernel Smoothing with Tensor Factor Model

In this section, we derive the kernel smoothing formula (**??**) under the vanilla tensor factor model. Under this setting, the relevant covariates $\mathbf{X}_1$ is still available for the 1-st mode and we would like to predict a new tensor $\mathcal{Y}^{new} \in \mathbb{R}^{I_1^{new} \times I_2 \times I_3}$ with new covariates $\mathbf{X}_1^{new} \in \mathbb{R}^{I_1^{new} \times D_1}$. However, we do not use the STEFA model to incorporate $\mathbf{X}_1$ in the model. Instead, we use an algorithm for solving noisy Tucker decomposition (**??**) and obtain an estimator of the signal part $\widehat{\mathcal{S}} = \widehat{\mathcal{F}} \times_1 \widehat{\mathbf{A}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3$. The informative covariates $\mathbf{X}_1$ and $\mathbf{X}_1^{new}$ are used non-parametrically.

Recall that we defined the kernel weight matrix $\mathbf{W} \in \mathbb{R}^{I_1^{new} \times I_1}$ with entry

$$w_{ij} = \frac{K_h(dist(\mathbf{x}_{1,i\cdot}^{new}, \mathbf{x}_{1,j\cdot}))}{\sum_{i=1}^{I_1} K_h(dist(\mathbf{x}_{1,i\cdot}^{new}, \mathbf{x}_{1,j\cdot}))}, \quad i \in [I_1^{new}] \text{ and } j \in [I_1].$$

where $K_h(\cdot)$ is the kernel function, $dist(\cdot,\cdot)$ is a pre-defined distance function such as the Euclidean distance, and $\mathbf{x}_{1,i\cdot}$ is the i-th row of $\mathbf{X}_1$.

For each row of $\mathbf{X}_1^{new}$, we will predict a tensor slice $\mathbf{Y}_i^{new} \in \mathbb{R}^{I_2 \times I_3}$. Let $\mathbf{y}_i^{new} = \text{vec}(\mathbf{Y}_i^{new})$, $\mathbf{Y} \triangleq \mathcal{M}_1(\mathcal{Y})^\top = [\mathbf{y}_1 \cdots \mathbf{y}_{I_1}]$, consisting of the signal part $\mathbf{S} \triangleq \mathcal{M}_1(\mathcal{S})^\top = [\mathbf{s}_1 \cdots \mathbf{s}_{I_1}]$ and the noise part $\mathbf{E} \triangleq \mathcal{M}_1(\mathcal{E})^\top = [\mathbf{e}_1 \cdots \mathbf{e}_{I_1}]$. Define $\mathbf{\Sigma}_y \triangleq \mathbb{E}\left[\mathbf{Y}^\top \mathbf{Y}\right]$ and $\mathbf{\Sigma}_i^{new} \triangleq \mathbb{E}\left[\mathbf{Y}^\top \mathbf{y}_i^{new}\right]$. The

best linear predictor for $\widehat{\mathbf{y}}_i^{new}$ based on $\mathbf{Y}$ is

$$\widehat{\mathbf{y}}_i^{new} = \mathbf{Y} \cdot \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Sigma}_i^{new}. \tag{21}$$

With knowledge of covariates $\mathbf{X}_1$, it is possible to estimate $\boldsymbol{\Sigma}_y^{-1}$ and $\boldsymbol{\Sigma}_i^{new}$ from $\mathbf{X}_1$ and $\mathbf{x}_{i\cdot}^{new}$. However, in practice it involves inverting a $I_1 \times I_1$ matrix $\boldsymbol{\Sigma}_y$ which may be computational costly when I_1 is large. The computational burden can be relieved by taking advantage of the Tucker low-rank structure.

To estimate $\boldsymbol{\Sigma}_i^{new}$, we note that $\boldsymbol{\Sigma}_i^{new} = \mathbb{E}\left[(I_2 I_3)^{-1} \mathbf{Y}^\top \mathbf{s}_i^{new}\right]$ where $\mathbf{s}_i^{new}$ is the signal part of $\mathbf{y}_i^{new}$. Thus, it can be estimated by $\widehat{\boldsymbol{\Sigma}}_i^{new} = (I_2 I_3)^{-1} \mathbf{Y}^\top \widehat{\mathbf{s}}_i^{new}$. We use kernel predictors for $\widehat{\mathbf{s}}_i^{new}$, that is,

$$\widehat{\mathbf{s}}_i^{new} = \frac{\sum_{j=1}^{I_1} \widehat{\mathbf{s}}_j K_h(dist(\mathbf{x}_{1,i\cdot}^{new}, \mathbf{x}_{1,j\cdot}))}{\sum_{i=1}^{I_1} K_h(dist(\mathbf{x}_{1,i\cdot}^{new}, \mathbf{x}_{1,j\cdot}))} = \sum_{j=1}^{I_1} w_{ij} \widehat{\mathbf{s}}_j. \tag{22}$$

With careful calculation, we have a simpler expression for $\widehat{\mathbf{y}}_i^{new}$. First, we have $\widehat{\mathbf{s}}_i^{new} = \widehat{\mathbf{S}} \mathbf{w}_{i\cdot}$ and

$$\widehat{\mathbf{y}}_i^{new} = \mathbf{Y} \widehat{\boldsymbol{\Sigma}}_y^{-1} \widehat{\boldsymbol{\Sigma}}_i^{new} = \mathbf{Y} \widehat{\boldsymbol{\Sigma}}_y^{-1} \cdot \mathbf{Y}^\top \widehat{\mathbf{S}} \mathbf{w}_{i\cdot} = \mathbf{Y} \widehat{\boldsymbol{\Sigma}}_y^{-1} \mathbf{Y}^\top \mathbf{Y} \widehat{\mathbf{A}}_1 \widehat{\mathbf{A}}_1^\top \mathbf{w}_{i\cdot} = \mathbf{Y} \widehat{\mathbf{A}}_1 \widehat{\mathbf{A}}_1^\top \mathbf{w} = \widehat{\mathbf{S}} \mathbf{w}_{i\cdot}. \tag{23}$$

Equation (23) shows that, under the tensor factor model, we do not need to actually calculate $\widehat{\boldsymbol{\Sigma}}_y^{-1}$ to obtain the best linear predictor (21). Kernel smoothing formula (**??**) is obtained by applying (23) to each i-th row of $\mathbf{X}_1^{new}$ and stacking the resulting tensor slices $\widehat{\mathbf{Y}}_i^{new}$ along the first mode for $i \in I_1^{new}$.

Appendix D More Simulation Results

D.1 Inequal Dimensions

In this section, we consider the setting where tensor $\mathcal{Y}$ has different dimensions, that is, I_1, I_2, I_3 are not equal. We fix $R = 3$, $I_1 = 100$ but vary α and $I_2, I_3 \geq I_1$. The ReMSE of estimating the loading matrices $\mathbf{A}_m$ and the tensor $\mathcal{Y}$ are reported in Table 2.

Although the dimensions for the three modes are artificially designed to be different in this simulation, no significant difference between $\ell_2(\widehat{\mathbf{A}}_m)$, $m \in [3]$ is observed. The error

in estimating the loading matrices of the three modes appears to be symmetric. With a fixed signal-to-noise ratio coefficient α and a fixed $I_{min} = I_1 = 100$, the performance of both projected Tucker and vanilla Tucker decomposition is not sensitive to the other two dimensions I_2, I_3.

(I_1, I_2, I_3)	R	α	IP-SVD				HOOI			
			$\ell_2(\widehat{\mathbf{A}}_1)$	$\ell_2(\widehat{\mathbf{A}}_2)$	$\ell_2(\widehat{\mathbf{A}}_3)$	$\text{ReMSE}_{\mathcal{Y}}$	$\ell_2(\widehat{\mathbf{A}}_1)$	$\ell_2(\widehat{\mathbf{A}}_2)$	$\ell_2(\widehat{\mathbf{A}}_3)$	$\text{ReMSE}_{\mathcal{Y}}$
(100,100,200)	3	0.3	0.805	0.805	0.820	0.885	1.703	1.703	1.718	3.647
			(0.260)	(0.242)	(0.253)	(0.283)	(0.014)	(0.016)	(0.007)	(0.782)
(100,100,400)	3	0.3	0.824	0.850	0.859	0.930	1.704	1.703	1.725	4.329
			(0.227)	(0.219)	(0.234)	(0.284)	(0.012)	(0.015)	(0.004)	(0.958)
(100,200,200)	3	0.3	0.840	0.828	0.782	0.903	1.706	1.718	1.719	4.072
			(0.223)	(0.213)	(0.208)	(0.264)	(0.014)	(0.006)	(0.006)	(0.802)
(100,200,400)	3	0.3	0.840	0.857	0.853	0.935	1.705	1.718	1.725	4.711
			(0.222)	(0.239)	(0.221)	(0.259)	(0.011)	(0.005)	(0.003)	(0.910)
(100,100,200)	3	0.5	0.264	0.278	0.274	0.279	1.641	1.635	1.655	1.715
			(0.073)	(0.076)	(0.067)	(0.071)	(0.172)	(0.177)	(0.167)	(0.348)
(100,100,400)	3	0.5	0.274	0.282	0.271	0.280	1.695	1.688	1.715	1.981
			(0.065)	(0.068)	(0.071)	(0.060)	(0.048)	(0.047)	(0.039)	(0.288)
(100,200,200)	3	0.5	0.258	0.277	0.262	0.268	1.677	1.686	1.685	1.825
			(0.061)	(0.062)	(0.068)	(0.063)	(0.095)	(0.117)	(0.124)	(0.323)
(100,200,400)	3	0.5	0.273	0.270	0.262	0.271	1.692	1.704	1.712	2.063
			(0.078)	(0.069)	(0.068)	(0.066)	(0.074)	(0.071)	(0.068)	(0.373)

Table 2: Unbalanced tensor dimensions. The average spectral and Frobenius Schatten q-$\sin\Theta$ loss ($q = 2$) for $\widehat{\mathbf{A}}_m$, $m \in [3]$ and average Frobenius loss for $\mathcal{Y}$ under various settings.

D.2 Comparison to the MMC Linear Tensor Regression

In this section, we compare our approach (IP-SVD) to the MMC tensor regression method of **?** on a linear tensor model. The $100 \times 100 \times 100$ observed tensor $\mathcal{Y}$ is generated in the same way as in Section **??** with a core tensor of $1 \times 1 \times 1$. That is we set $I_1 = I_2 = I_3 = 100$ and $R_1 = R_2 = R_3 = 1$. Covariates $\mathbf{X}_m, m = 1, 2, 3$ of 100×1 are randomly sampled from a uniform distribution on $[0, 1]$ i.i.d.. The parametric loading matrix $\mathbf{G}_m$ is quadratic with respect to $\mathbf{X}_m$ such that $[\mathbf{G}_m]_i \propto 1 + [\mathbf{X}_m]_i + [\mathbf{X}_m]_i^2$. The non-parametric loading matrix $\mathbf{\Gamma}_m$ is added in a similar way to Section **??**. In summary, the observed tensor is generated according to

$$\mathcal{Y} = \mathcal{F} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times \mathbf{A}_3 + \mathcal{E}, \tag{24}$$

where

$$\mathbf{A}_m = \mathbf{G}_m + \mu\mathbf{\Gamma}_m,$$
$$\mathbf{G}_m = \mathbf{Z}_m/\|\mathbf{Z}_m\|,$$
$$\mathbf{Z}_m = 1 + \mathbf{X}_m + \mathbf{X}_m^2,$$

and $\mathcal{E}$ is a $I_1 \times I_2 \times I_3$ tensor with i.i.d. standard Gaussian entries, $\mathcal{F}$ is a $1 \times 1 \times 1$ (a scalar) of value I^α and $\mathbf{X}_m$ has i.i.d. Uniform(0,1) entries. Again, we use α to control the signal-to-noise ratio and use μ to control the relative strength of non-parametric loading parts.

Three models are used to fit $\mathcal{Y}$. IP-SVD(NP) model denotes the IP-SVD approach that we fit $\mathcal{Y}$ with correctly specified ranks and sieve orders and with the *non-parametric (NP)* loading parts as in (24). IP-SVD(P) model is a similar model of IP-SVD(NP) except that we ignore the non-parametirc part $\mathbf{\Gamma}$ and only fits the *parametric (P)* part $\mathbf{A}_m = \mathbf{G}_m$. MMC-LTR model stands for the multiple-mode-covariate linear regression model from **?** where each $\mathbf{A}_m$ is assumed to be linear in $\mathbf{X}_m$ (and therefore, misspecifies the model with low sieve ranks). We report the relative MSE (ReMSE, averaged over 100 repetitions) of the three methods for different signal-to-noise ratios (varying α) and for different non-parametric components (varying μ) in Table 3.

α	2			1		
μ	1	0.1	0	1	0.1	0
IP-SVD(NP)	0.004 (1.5e-4)	0.002 (6.7e-5)	0.002 (6.6e-5)	0.343 (0.019)	0.174 (0.007)	0.172 (0.007)
IP-SVD(P)	0.931 (0.002)	0.170 (0.001)	2.6e-4 (6.7e-5)	0.932 (0.002)	0.172 (0.001)	0.026 (0.007)
MMC-LTR	0.937 (0.008)	0.271 (0.112)	0.199 (0.138)	0.937 (0.008)	0.272 (0.112)	0.201 (0.138)

Table 3: Mean and standard deviation of the Relative MSE for the three approaches under different signal-to-noise ratios and different strength of non-parametric parts in 100 repetitions.

When the relative strength of the non-parametric loading $\mathbf{\Gamma}_m$ is strong ($\mu = 1$), we have IP-SVD(NP) > IP-SVD(P) > MMC-LTR, where > means "performs better than", for both

settings of moderate and strong signal-to-noise ratio. When the relative strength of the non-parametric loading $\mathbf{\Gamma}_m$ is weak ($\mu = 0.1$) or non-existing ($\mu = 0$), the IP-SVD still performs better than MMC-LTR. The IP-SVD(NP) has a disadvantage relative to IP-SVD(P) under this setting, especially when the model is fully parametric ($\mu = 0$). However, when the signal-to-noise ratio is strong $\alpha = 2$, IP-SVD(NP) still performs the best in face of the weak relative strength of the non-parametric loading $\mathbf{\Gamma}_m$ ($\mu = 0.1$).

We conclude at least for the specific setting with linear linkage function, nonlinear loading factors and non-parametric loading parts, IP-SVD outperforms the tensor regression method due to IP-SVD's capability as a semiparametric model with sieve expansions. At the same time, we acknowledge that the tensor regression method can handle more complicated linkage functions such as logistic model and probit model. In general, the STEFA model as a unsupervised method is a complement to the supervised MMC tensor regression model (**?**).

Appendix E Proofs of Technical Lemmas

Proof of Lemma 1. By the definitions of $\mathbf{A}_1$ and $\mathbf{Z}_1$, we write

$$\mathbf{A}_1\mathbf{Z}_1^\top = \mathbf{P}_1\mathbf{G}_1\mathcal{M}_1(\mathcal{F})\big((\mathbf{P}_2\mathbf{G}_2)\otimes(\mathbf{P}_3\mathbf{G}_3)\big)^\top(\mathbf{U}_2\otimes\mathbf{U}_3)\big((\mathbf{U}_2\otimes\mathbf{U}_3)^\top\mathcal{M}_1^\top(\mathcal{E})\mathbf{U}_1\mathbf{U}_1^\top\big) \quad (25)$$

where $\mathbf{U}_m\mathbf{U}_m^\top = \mathbf{P}_m$, $\mathbf{U}_m \in \mathbb{R}^{I_m\times J_m}$ and $\mathbf{U}_m^\top\mathbf{U}_m = \mathbf{I}_{J_m}$. It suffices to prove the upper bound of $\|\mathbf{B}_1(\mathbf{U}_2\otimes\mathbf{U}_3)^\top\mathcal{M}_1^\top(\mathcal{E})\mathbf{U}_1\|$ where $\mathbf{B}_1 = \mathbf{A}_1(\mathbf{U}_2\otimes\mathbf{U}_3)$ is an $I_1\times(J_2J_3)$ deterministic matrix.

Denote $\mathbf{E}_1 = \mathcal{M}_1^\top(\mathcal{E})\mathbf{U}_1 \in \mathbb{R}^{(I_2I_3)\times J_1}$. By Assumption **??**, $\mathbf{E}_1 = (\mathbf{e}_{1,1},\cdots,\mathbf{e}_{1,I_2I_3})^\top$ has i.i.d. rows and each row is a J_1-dimensional centered sub-exponential random vector in that

$$\sup_{\|\mathbf{v}\|\le 1}\mathbb{P}\big(|\langle\mathbf{v},\mathbf{e}_{1,j}\rangle|\ge t\big)\le\exp(-Ct)\quad\text{for all } j\in[I_2I_3] \quad (26)$$

for any $t > 1$. Meanwhile, $\mathbb{E}(\mathbf{e}_{1,j}\mathbf{e}_{1,j}^\top) = \mathbf{I}_{J_1}$ for all $j \in [I_2I_3]$. By (26), there exists an absolute

constant $C_1 > 0$ so that

$$\mathbb{P}\Big(\max_{j\in[I_2I_3],k\in[J_1]} \big|e_{1,j}(k)\big| \geq C_1 \log(I_2I_3J_1)\Big) \leq \frac{1}{(I_2I_3J_1)^4}. \tag{27}$$

Denote the above event by $\mathfrak{E}_0$. We obtain $\max_j \|\mathbf{e}_{1,j}\| \leq C_1\sqrt{J_1}\log(I_2I_3J_1)$ on $\mathfrak{E}_0$. Now, we denote $\delta_1 = C_1\sqrt{J_1}\log(I_2I_3J_1)$.

Denote $\{\widetilde{\mathbf{u}}_{23,j}\}_{j=1}^{I_2I_3}$ the columns of $(\mathbf{U}_2 \otimes \mathbf{U}_3)^\top$. We write

$$\begin{aligned}\mathbf{B}_1(\mathbf{U}_2 \otimes \mathbf{U}_3)^\top \mathbf{E}_1 =& \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top \\ =& \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| \leq \delta_1) + \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| > \delta_1),\end{aligned}$$

where the second term on the RHS is simply 0 on event $\mathfrak{E}_0$. It suffices to bound the first term, which is a sum of independent random matrices. Write

$$\begin{aligned}\Big\| \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| \leq \delta_1)\Big\| \leq& \Big\| \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\Big(\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| \leq \delta_1) - \mathbb{E}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| \leq \delta_1)\Big)\Big\| \\ &+ \Big\| \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbb{E}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| \leq \delta_1)\Big\|. \end{aligned} \tag{28}$$

Since $\mathbb{E}\mathbf{e}_{1,j} = 0$ and $\|\mathbf{B}_1\| \leq \|\mathbf{A}_1\| \leq \kappa_0\lambda_{\min}\sqrt{I_1I_2I_3}$, together with (27), we get

$$\begin{aligned}\Big\| \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbb{E}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| \leq \delta_1)\Big\| =& \Big\| \sum_{j=1}^{I_2I_3} \mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbb{E}\mathbf{e}_{1,j}^\top \mathbb{1}(\|\mathbf{e}_{1,j}\| > \delta_1)\Big\| \\ \leq& \|\mathbf{B}_1\| \cdot \sum_{j=1}^{I_2I_3} \mathbb{E}\|\mathbf{e}_{1,j}\|\mathbb{1}(\|\mathbf{e}_{1,j}\| > \delta_1) = I_2I_3\|\mathbf{B}_1\| \cdot \mathbb{E}\|\mathbf{e}_{1,1}\|\mathbb{1}(\|\mathbf{e}_{1,1}\| > \delta_1) \\ \leq& I_2I_3\|\mathbf{B}_1\| \cdot \mathbb{E}^{1/2}\big(\|\mathbf{e}_{1,1}\|^2\big) \cdot \mathbb{P}^{1/2}\big((\|\mathbf{e}_{1,1}\| > \delta_1\big) \\ \leq& I_2I_3\kappa_0\lambda_{\min}\sqrt{I_1I_2I_3} \cdot C_2\sqrt{J_1} \cdot \frac{1}{(I_2I_3J_1)^2} \leq C_2\frac{\kappa_0\lambda_{\min}\sqrt{I_1I_2I_3}}{I_2I_3J_1}.\end{aligned}$$

Now it suffices to prove the upper bound of first term in RHS of (28), which is the spectral

norm of the sum of independent random matrices. By definition,

$$\Big\|\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\Big(\mathbf{e}_{1,j}^\top\mathbb{1}(\|\mathbf{e}_{1,j}\|\le\delta_1)-\mathbb{E}\mathbf{e}_{1,j}^\top\mathbb{1}(\|\mathbf{e}_{1,j}\|\le\delta_1)\Big)\Big\|\le 2\delta_1\kappa_0\lambda_{\min}\sqrt{I_1I_2I_3},\quad \forall j\in[I_2I_3]$$

and

$$\begin{aligned}
&\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\cdot\mathbb{1}(\|\mathbf{e}_{1,j}\|\le\delta_1)\Big\|\\
&\quad\le\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\Big\|+\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\cdot\mathbb{1}(\|\mathbf{e}_{1,j}\|>\delta_1)\Big\|\\
&\quad\le J_1I_1I_2I_3\kappa_0^2\lambda_{\min}^2+\sum_{j=1}^{I_2I_3}\|\widetilde{\mathbf{u}}_{23,j}\|^2\kappa_0^2\lambda_{\min}^2I_1I_2I_3\cdot\mathbb{E}\|\mathbf{e}_{1,j}\|^2\mathbb{1}(\|\mathbf{e}_{1,j}\|>\delta_1)\\
&\quad\le J_1\kappa_0^2\lambda_{\min}^2I_1I_2I_3+J_2J_3\kappa_0^2\lambda_{\min}^2I_1I_2I_3\cdot C_2J_1\cdot\frac{1}{(I_2I_3J_1)^2}\le 2J_1\kappa_0^2\lambda_{\min}^2I_1I_2I_3,
\end{aligned}$$

where the last inequality holds since $I_m\ge J_m$. Similarly, we have

$$\begin{aligned}
&\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\cdot\mathbb{1}(\|\mathbf{e}_{1,j}\|\le\delta_1)\Big\|\\
&\quad\le\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\Big\|+\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\cdot\mathbb{1}(\|\mathbf{e}_{1,j}\|>\delta_1)\Big\|\\
&\quad\le\|\mathbf{B}_1\|_{\mathrm{F}}^2+\Big\|\sum_{j=1}^{I_2I_3}\mathbb{E}\mathbf{e}_{1,j}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbf{e}_{1,j}^\top\cdot\mathbb{1}(\|\mathbf{e}_{1,j}\|>\delta_1)\Big\|\\
&\quad\le R_1\kappa_0^2\lambda_{\min}^2I_1I_2I_3+\sum_{j=1}^{I_2I_3}\widetilde{\mathbf{u}}_{23,j}^\top\mathbf{B}_1^\top\mathbf{B}_1\widetilde{\mathbf{u}}_{23,j}\mathbb{E}\langle\mathbf{e}_{1,1},\mathbf{v}\rangle^2\cdot\mathbb{1}(\|\mathbf{e}_{1,1}\|>\delta_1)\big]\le 2R_1\kappa_0^2\lambda_{\min}^2I_1I_2I_3,
\end{aligned}$$

where $\mathbf{v}$ is any fixed vector in $\mathbb{R}^{J_1}$ with unit norm.

Then, by matrix Bernstein inequality (**?**), with probability at least $1-I_1^{-2}$,

$$\begin{aligned}
&\Big\|\sum_{j=1}^{I_1}\mathbf{B}_1\Big(\mathbf{e}_{1,j}\mathbb{1}(\|\mathbf{e}_{1,j}\|\le\delta_1)-\mathbb{E}\mathbf{e}_{1,j}\mathbb{1}(\|\mathbf{e}_{1,j}\|\le\delta_1)\Big)\widetilde{\mathbf{u}}_j^\top\Big\|\\
&\qquad\qquad\le C_1\kappa_0\lambda_{\min}(I_1I_2I_3)^{1/2}\big(\sqrt{J_1\log I_1}+\delta_1\log I_1\big),
\end{aligned}$$

where we assumed $R_m\le J_m$. Since $\delta_1\asymp\sqrt{J_1}\log(I_1)$ and $I_1>J_1$, we get with probability

at least $1-2I_1^{-2}$ that

$$\left\|\mathbf{B}_1(\mathbf{U}_2\otimes\mathbf{U}_3)^\top\mathbf{E}_1\mathbf{U}_1\mathbf{U}_1^\top\right\|\le C_3\kappa_0\lambda_{\min}(I_1I_2I_3)^{1/2}\cdot\sqrt{J_1}\log^2 I_1$$

for some absolute constant $C_3>0$.

Therefore, by (25), we get with probability at least $1-2I_1^{-2}$ that

$$\|\mathbf{A}_1\mathbf{Z}_1^\top\|\le C_4\kappa_0\lambda_{\min}(I_1I_2I_3)^{1/2}\cdot\sqrt{J_1}\log^2 I_1$$

for some absolute constant $C_4>0$, which proves Lemma 1. □

Proof of Lemma 2. Similarly as the proof of Lemma 1, we denote $\mathbf{E}_2=\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)$ the $I_1\times(J_2J_3)$ matrix with independent rows $\left(\mathbf{e}_{2,i}^\top\right)_{i=1}^{I_1}$. Here, $\mathbf{e}_{2,i}$ is a sub-exponential random vector with $\mathbb{E}\mathbf{e}_{2,i}\mathbf{e}_{2,i}^\top=\mathbf{I}_{J_2J_3}$ and $\mathbb{E}\mathbf{e}_{2,i}=\mathbf{0}$.

Similarly, we denote $\{\widetilde{\mathbf{u}}_i\}_{i=1}^{I_1}$ the columns of $\mathbf{U}_1^\top$. Then, we write

$$\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)=\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top$$

and as a result

$$\begin{aligned}
&\left(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\right)\left(\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{U}_2\otimes\mathbf{U}_3)\right)^\top-J_2J_3\mathbf{I}_{J_1}\\
=&\Big(\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top\Big)\Big(\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top\Big)^\top-J_2J_3\mathbf{I}_{J_1}\\
=&\Big(\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top\mathbf{e}_{2,i}\widetilde{\mathbf{u}}_i^\top-J_2J_3\mathbf{I}_{J_1}\Big)+\sum_{1\le i_1\ne i_2\le I_1}\widetilde{\mathbf{u}}_{i_1}\mathbf{e}_{2,i_1}^\top\mathbf{e}_{2,i_2}\widetilde{\mathbf{u}}_{i_2}^\top.
\end{aligned}\tag{29}$$

Observe that $\mathbb{E}\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top\mathbf{e}_{2,i}\widetilde{\mathbf{u}}_i^\top=J_2J_3\mathbf{I}_{J_1}$. Note that

$$\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top\mathbf{e}_{2,i}\widetilde{\mathbf{u}}_i^\top-J_2J_3\mathbf{I}_{J_1}=\sum_{i=1}^{I_1}\|\mathbf{e}_{2,i}\|^2\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i^\top-J_2J_3\mathbf{I}_{J_1}.$$

Similarly as proof of Lemma 1, denote $\delta_1=C_1\sqrt{J_2J_3}\log I_1$ so that $\mathbb{P}(\max_i\|\mathbf{e}_{2,i}\|\ge\delta_1)\le$

$(I_1J_2J_3)^{-3}$. Following the same treatment there, we can show that

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{2,i}^\top\mathbf{e}_{2,i}\widetilde{\mathbf{u}}_i^\top - J_2J_3\mathbf{I}_{J_1}\Big\| \geq (J_2J_3)^{1/2}\big(C_2\sqrt{J_1\log I_1}+C_3\log^2 I_1\big)\Big) \leq I_1^{-2} \tag{30}$$

with $C_2, C_3 > 0$ being absolute constants, where we used the facts (needed in matrix Bernstein inequality), with $\mathbf{e}_1$ being the first column of $\mathcal{M}_1(\mathcal{E})^\top$,

$$\begin{aligned}
\Big\|\sum_{i=1}^{I_1}\mathbb{E}\big(\|\mathbf{e}_{2,i}\|^2-J_2J_3\big)^2\|\widetilde{\mathbf{u}}_i\|^2\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i^\top\Big\| &= \mathbb{E}\big(\|\mathbf{e}_{2,1}\|^2-J_2J_3\big)^2\Big\|\sum_{i=1}^{I_1}\|\widetilde{\mathbf{u}}_i\|^2\widetilde{\mathbf{u}}_i\widetilde{\mathbf{u}}_i^\top\Big\| \\
&\leq \mathbb{E}\big(\|\mathbf{e}_{2,1}\|^2-J_2J_3\big)^2 = \mathbb{E}\|\mathbf{e}_{2,1}\|^4-(J_2J_3)^2 \overset{\mathbf{U}_{23}=\mathbf{U}_2\otimes\mathbf{U}_3}{=} \mathbb{E}\big(\mathbf{e}_1^\top\mathbf{U}_{23}\mathbf{U}_{23}^\top\mathbf{e}_1\big)^2-(J_2J_3)^2 \\
\overset{\mathbf{U}_{23}=(\mathbf{u}_{23,j})_{j=1}^{J_2J_3}}{=} &\mathbb{E}\Big(\sum_{j=1}^{J_2J_3}\langle\mathbf{e}_1,\mathbf{u}_{23,j}\rangle^2\Big)^2-(J_2J_3)^2 \\
&= \sum_{j=1}^{J_2J_3}\mathbb{E}\langle\mathbf{e}_1,\mathbf{u}_{23,j}\rangle^4 + \sum_{1\leq j\neq j'\leq J_2J_3}\mathbb{E}\langle\mathbf{e}_1,\mathbf{u}_{23,j}\rangle^2\langle\mathbf{e}_1,\mathbf{u}_{23,j'}\rangle^2-(J_2J_3)^2 \\
&= \sum_{j=1}^{J_2J_3}\mathbb{E}\langle\mathbf{e}_1,\mathbf{u}_{23,j}\rangle^4 + \sum_{1\leq j\neq j'\leq J_2J_3}\|\mathbf{u}_{23,j}\|^2\|\mathbf{u}_{23,j'}\|^2-(J_2J_3)^2 \\
&= \sum_{j=1}^{J_2J_3}\Big(\mathbb{E}\langle\mathbf{e}_1,\mathbf{u}_{23,j}\rangle^4-\|\mathbf{u}_{23,j}\|^4\Big)+\|\mathbf{U}_{23}\|_{\mathrm{F}}^2\|\mathbf{U}_{23}\|_{\mathrm{F}}^2-(J_2J_3)^2 \\
&= \sum_{j=1}^{J_2J_3}\Big(\mathbb{E}\langle\mathbf{e}_1,\mathbf{u}_{23,j}\rangle^4-\|\mathbf{u}_{23,j}\|^4\Big) = O(J_2J_3).
\end{aligned}$$

We now deal with the second term in RHS of (29). Observe that

$$g(\mathbf{E}_2) = \sum_{1\leq i_1\neq i_2\leq I_1}\widetilde{\mathbf{u}}_{i_1}\widetilde{\mathbf{u}}_{i_2}^\top\langle\mathbf{e}_{2,i_1},\mathbf{e}_{2,i_2}\rangle$$

is a generalized U-statistic. Let $\{\widetilde{\mathbf{e}}_{2,i}\}_{i=1}^{I_1}$ be an independent copy of $\{\mathbf{e}_{2,i}\}_{i=1}^{I_1}$. By the tail probability of decoupling of U-statistics ((**?**, Theorem 3.4.1)), for all $t>0$,

$$\mathbb{P}\Big(\Big\|\sum_{1\leq i_1\neq i_2\leq I_1}\widetilde{\mathbf{u}}_{i_1}\widetilde{\mathbf{u}}_{i_2}^\top\langle\mathbf{e}_{2,i_1},\mathbf{e}_{2,i_2}\rangle\Big\|\geq t\Big) \leq C_1\cdot\mathbb{P}\Big(\Big\|\sum_{1\leq i_1\neq i_2\leq I_1}\widetilde{\mathbf{u}}_{i_1}\widetilde{\mathbf{u}}_{i_2}^\top\langle\mathbf{e}_{2,i_1},\widetilde{\mathbf{e}}_{2,i_2}\rangle\Big\|\geq C_2t\Big) \tag{31}$$

for some absolute constants $C_1, C_2 > 0$. Clearly,

$$\Big\| \sum_{1\le i_1\neq i_2\le I_1} \widetilde{\mathbf{u}}_{i_1}\widetilde{\mathbf{u}}_{i_2}^\top \langle \mathbf{e}_{2,i_1}, \widetilde{\mathbf{e}}_{2,i_2}\rangle \Big\| \le \big\| \big(\mathbf{U}_1^\top \mathbf{E}_2\big)\big(\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2\big)^\top \big\| + \Big\| \sum_{i=1}^{I_1} \widetilde{\mathbf{u}}_i \widetilde{\mathbf{u}}_i^\top \langle \mathbf{e}_{2,i}, \widetilde{\mathbf{e}}_{2,i}\rangle \Big\|. \tag{32}$$

By a similar treatment as in the proof of Lemma 1, we get

$$\mathbb{P}\big(\|\mathbf{U}_1^\top \mathbf{E}_2\| \ge C_3\sqrt{J_1 \log I_1} + C_4\sqrt{J_2 J_3}\log^2 I_1\big) \le I_1^{-2}.$$

Denote $\mathfrak{E}_1$ the above event. To get a sharp upper bound for $\|(\mathbf{U}_1^\top \mathbf{E}_2)(\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2)^\top\|$, we fix $\widetilde{\mathbf{E}_2}$ and recall the definition $\mathbf{U}_1^\top \mathbf{E}_2 = \mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{U}_2 \otimes \mathbf{U}_3)$. Define $\mathbf{E}_1 = \mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E}) \in \mathbb{R}^{J_1 \times (I_2 I_3)}$, which has i.i.d. columns $\{\mathbf{e}_{1,i}\}_{i=1}^{I_2 I_3}$. Denote $\{\widetilde{\mathbf{u}}_{23,i}^\top\}_{i=1}^{I_2 I_3}$ the rows of $\mathbf{U}_2 \otimes \mathbf{U}_3$. Then, we write

$$(\mathbf{U}_1^\top \mathbf{E}_2)(\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2)^\top = \sum_{i=1}^{I_2 I_3} \mathbf{e}_{1,i}\widetilde{\mathbf{u}}_{23,i}^\top (\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2)^\top.$$

Similarly by the treatment of the proof of Lemma 1, conditioned on $\widetilde{\mathbf{E}}_2$, we have

$$\mathbb{P}\big(\|(\mathbf{U}_1^\top \mathbf{E}_2)(\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2)^\top\| \ge \|\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2\| \cdot \big(C_3\sqrt{J_1 \log I_1} + C_4\sqrt{J_1}\log^2 I_1\big)\big|\widetilde{\mathbf{E}}_2\big) \le I_1^{-2}.$$

Together with the event $\mathfrak{E}_1$, we conclude that with probability at least $1 - 2I_1^{-2}$,

$$\|(\mathbf{U}_1^\top \mathbf{E}_2)(\mathbf{U}_1^\top \widetilde{\mathbf{E}}_2)^\top\| \le C_3' J_1 \log^{5/2} I_1 + C_4'\sqrt{J_1 J_2 J_3}\log^4 I_1 \tag{33}$$

for some absolute constants $C_3', C_4' > 0$.

For the second term in (32), we still apply the truncation treatment as in the proof of Lemma 1. In this case, note that there exists an event $\mathfrak{E}_2$ with $\mathbb{P}(\mathfrak{E}_2) \ge 1 - I_1^{-2}$ such that $\max_{i\in[I_1]} \|\widetilde{\mathbf{e}}_{2,i}\| \le C_0\sqrt{J_2 J_3}\log I_1$. Conditioned on $\widetilde{\mathbf{e}}_{2,i}$, we have $\mathbb{P}(|\langle \mathbf{e}_{2,i}, \widetilde{\mathbf{e}}_{2,i}\rangle| \ge C_1'\|\widetilde{\mathbf{e}}_{2,i}\|\log I_1) \le I_1^{-4}$ for some absolute constant $C_1' > 0$. By a similar proof, we can obtain

$$\mathbb{P}\Big(\Big\| \sum_{i=1}^{I_1} \widetilde{\mathbf{u}}_i \widetilde{\mathbf{u}}_i^\top \langle \mathbf{e}_{2,i}, \widetilde{\mathbf{e}}_{2,i}\rangle \Big\| \ge C_3\sqrt{J_2 J_3 \log I_1} + C_4\sqrt{J_2 J_3}\log^3 I_1\Big) \le 2I_1^{-2}. \tag{34}$$

Putting (34), (33), (32), (31) together, we get

$$\mathbb{P}\Big(\Big\| \sum_{1\le i_1\neq i_2\le I_1} \widetilde{\mathbf{u}}_{i_1}\widetilde{\mathbf{u}}_{i_2}^\top \langle \mathbf{e}_{2,i_1}, \mathbf{e}_{2,i_2}\rangle \Big\| \ge C_3' J_1 \log^{5/2} I_1 + C_4'\sqrt{J_1 J_2 J_3}\log^4 I_1\Big) \le 4I_1^{-2}.$$

Then, together with (30) and (29), we get with probability at least $1 - 5I_1^{-2}$ that

$$\|\mathbf{Z}_1\mathbf{Z}_1^\top - J_2J_3\mathbf{P}_1\| \le C_3\sqrt{J_1J_2J_3}\log^4 I_1 + C_4J_1\log^{5/2} I_1,$$

which proves Lemma 2. □

Proof of Lemma 3. The strategy is similar to that in the proof of Lemma 1. Denote $\mathbf{E}_1 = \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{G}_3)/\sqrt{I_2I_3}$ the $I_1 \times (R_2R_3)$ random matrices with i.i.d rows $\{\mathbf{e}_{1,i}^\top\}_{i=1}^{I_1}$ satisfying $\mathbb{E}\mathbf{e}_{1,i} = \mathbf{0}$ and $\mathbb{E}\mathbf{e}_{1,i}\mathbf{e}_{1,i}^\top = \mathbf{I}_{R_2R_3}$. Then,

$$\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{G}_3) = \sqrt{I_2I_3}\mathbf{U}_1^\top\mathbf{E}_1 = \sqrt{I_2I_3}\cdot\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{1,i}^\top,$$

where $\{\widetilde{\mathbf{u}}_i^\top\}_{i=1}^{I_1}$ denotes the rows of $\mathbf{U}_1$. By a similar treatment, we have

$$\mathbb{P}\Big(\max_{1\le i\le I_1}\|\mathbf{e}_{1,i}\|^2 \ge C_1R_2R_3\log^2 I_1\Big) \le I_1^{-4}.$$

Denote the above event by $\mathfrak{E}_0$. Define $\delta_1 = C_1'\sqrt{R_2R_3}\log I_1$ so that $\mathbb{P}(\max_i\|\mathbf{e}_{1,i}\| \ge \delta_1) \le I_1^{-4}$. Then, we write

$$\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{G}_3) = \sqrt{I_2I_3}\cdot\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{1,i}^\top\mathbb{1}(\|\mathbf{e}_{1,i}\| \le \delta_1) + \sqrt{I_2I_3}\cdot\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{1,i}^\top\mathbb{1}(\|\mathbf{e}_{1,i}\| > \delta_1). \tag{35}$$

The second term in RHS of (35) is simply 0 on event $\mathfrak{E}_0$. It suffices to investigate the first term on RHS of (35). We will apply the matrix Bernstein inequality as in the proof of Lemma 1. Indeed, we can show that

$$\mathbb{P}\bigg(\Big\|\sum_{i=1}^{I_1}\widetilde{\mathbf{u}}_i\mathbf{e}_{1,i}^\top\mathbb{1}(\|\mathbf{e}_{1,i}\| \le \delta_1)\Big\| \ge C_1\sqrt{J_1\log I_1} + C_2\sqrt{R_2R_3}\log^2 I_1\bigg) \le I_1^{-2}$$

for some absolute constants $C_1, C_2 > 0$. Since the proof is identical to the proof of Lemma 1,

we skip it here. □

Proof of Lemma 4. We only prove (11) since the proof of (12) is similar. We begin with a standard discretization step, see (**?**, Lemma 5). There exists a subset $\mathfrak{D}_{1/3} \subset \mathfrak{B}(J_3, R_3)$ such that $\log \mathrm{Card}(\mathfrak{D}_{1/3}) \leq c_1 J_3 R_3$ for some absolute constant $c_1 > 0$ and for any $\mathbf{B} \in \mathfrak{B}(J_3, R_3)$,

$$\min_{\mathbf{D} \in \mathfrak{D}_{1/3}} \|\mathbf{B} - \mathbf{D}\|_{\mathrm{F}} \leq 1/3.$$

It is easy to show that

$$\sup_{\mathbf{A} \in \mathfrak{B}(J_3, R_3)} \big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{A})\big\| / \sqrt{I_2} \leq \frac{3}{2} \cdot \max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D})\big\| / \sqrt{I_2}. \tag{36}$$

It suffices to prove the upper bound in RHS of (36). Under Assumption **??**, there exists an event $\mathfrak{E}_0$ with $\mathbb{P}(\mathfrak{E}_0) \geq 1 - I_1^{-5}$ in which

$$\max_{\omega \in [I_1] \times [I_2] \times [I_3]} |e_\omega| \leq C_0 \log I_1$$

for some absolute constant $C_0 > 0$. Denote $\delta_0 = C_0 \log I_1$ and $\|\mathcal{E}\|_\infty = \max_\omega |e_\omega|$. We write

$$\begin{aligned} \max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D})\big\| / \sqrt{I_2} &\leq \max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D}) \mathbb{1}(\|\mathcal{E}\|_\infty \leq \delta_0)\big\| / \sqrt{I_2} \\ &\quad + \max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D}) \mathbb{1}(\|\mathcal{E}\|_\infty > \delta_0)\big\| / \sqrt{I_2}. \end{aligned} \tag{37}$$

Conditioned on event $\mathfrak{E}_0$, the second term in RHS of (37) is simply 0. It suffices to prove the upper bound of first term in RHS of (37). Write

$$\begin{aligned} &\max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathcal{M}_1(\mathcal{E})(\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D}) \mathbb{1}(\|\mathcal{E}\|_\infty \leq \delta_0)\big\| / \sqrt{I_2} \\ &\qquad \leq \max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathbb{E}\big[\mathcal{M}_1(\mathcal{E}) \mathbb{1}(\|\mathcal{E}\|_\infty \leq \delta_0)\big](\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D})\big\| / \sqrt{I_2} \\ &\quad + \max_{\mathbf{D} \in \mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \Big[\mathcal{M}_1(\mathcal{E}) \mathbb{1}(\|\mathcal{E}\|_\infty \leq \delta_0) - \mathbb{E}\mathcal{M}_1(\mathcal{E}) \mathbb{1}(\|\mathcal{E}\|_\infty \leq \delta_0)\Big](\mathbf{G}_2 \otimes \mathbf{U}_3 \mathbf{D})\big\| / \sqrt{I_2}. \end{aligned} \tag{38}$$

To upper bound the first term in RHS of (38), note that $\forall \omega \in [I_1] \times [I_2] \times [I_3]$,

$$\big|\mathbb{E} e_\omega \mathbb{1}(|e_\omega| \le \delta_0)\big| = \big|\mathbb{E} e_\omega \mathbb{1}(|e_\omega| > \delta_0)\big| \overset{\text{Cauchy-Scwharz}}{\le} \mathbb{P}^{1/2}(|e_\omega| > \delta_0).$$

Then, the first term in RHS of (38) is bounded as

$$\begin{aligned}
&\max_{\mathbf{D}\in\mathfrak{D}_{1/3}} \big\|\mathbf{U}_1^\top \mathbb{E}\big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0)\big](\mathbf{G}_2 \otimes \mathbf{U}_3\mathbf{D})\big\|/\sqrt{I_2} \\
&\quad \le \|\mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0)\|_{\mathrm{F}} = \left(\sum_{\omega\in[I_1]\times[I_2]\times[I_3]} \mathbb{P}(|e_\omega| > \delta_0)\right)^{1/2} \le (I_1 I_2 I_3 I_1^{-5})^{1/2} = O(I_1^{-1}).
\end{aligned} \tag{39}$$

We now continue the upper bound the second term in RHS of (38). Similarly, let $\mathfrak{R}_{1/3}(J_1)$ and $\mathfrak{R}_{1/3}(R_2R_3)$ the 1/3-net of $\mathfrak{B}(J_1, 1)$ and $\mathfrak{B}(R_2R_3, 1)$, respectively. Then, given any vector $\mathbf{u} \in \mathfrak{B}(J_1, 1)$ and $\mathbf{w} \in \mathfrak{B}(R_2R_3, 1)$, we have

$$\min_{\mathbf{v}\in\mathfrak{R}_{1/3}(J_1)} \|\mathbf{u} - \mathbf{v}\| \le 1/3 \quad \text{and} \quad \min_{\mathbf{v}\in\mathfrak{R}_{1/3}(R_2R_3)} \|\mathbf{v} - \mathbf{w}\| \le 1/3$$

where $\|\cdot\|$ represents ℓ_2-norm for vectors. Meanwhile, $\log \mathrm{Card}(\mathfrak{R}_{1/3}(J_1)) \le c_1 J_1$ and $\log \mathrm{Card}(\mathfrak{R}_{1/3}(R_2R_3)) \le c_2 R_2 R_3$ for some absolute constants $c_1, c_2 > 0$. Then,

$$\begin{aligned}
&\max_{\mathbf{D}\in\mathfrak{D}_{1/3}} \Big\|\mathbf{U}_1^\top \Big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0) - \mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0)\Big](\mathbf{G}_2 \otimes \mathbf{U}_3\mathbf{D})\Big\|/\sqrt{I_2} \\
&\quad \le \frac{9}{2} \max_{\mathbf{D}\in\mathfrak{D}_{1/3}} \max_{\substack{\mathbf{u}\in\mathfrak{R}_{1/3}(J_1) \\ \mathbf{w}\in\mathfrak{R}_{1/3}(R_2R_3)}} \mathbf{u}^\top \mathbf{U}_1^\top \Big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0) - \mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0)\Big](\mathbf{G}_2 \otimes \mathbf{U}_3\mathbf{D})\mathbf{w}/\sqrt{I_2}.
\end{aligned}$$

Now, we fix $\mathbf{u}, \mathbf{w}$ and $\mathbf{D}$, and prove a concentration inequality for the above RHS, then we finish the proof by a union bound. Denote $\widetilde{\mathbf{u}} = \mathbf{U}_1\mathbf{u} \in \mathbb{R}^{I_1}$ and $\widetilde{\mathbf{v}} = (\mathbf{G}_2 \otimes \mathbf{U}_3\mathbf{D})\mathbf{w}/\sqrt{I_2} \in \mathbb{R}^{I_2 I_3}$. Clearly, we have $\max\big\{\|\widetilde{\mathbf{u}}\|, \|\widetilde{\mathbf{v}}\|\big\} \le 1$. Now, we write

$$\begin{aligned}
&\widetilde{\mathbf{u}}^\top \Big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0) - \mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty \le \delta_0)\Big]\widetilde{\mathbf{v}} \\
&\qquad = \sum_{\omega=(\omega_1,\omega_2,\omega_3)\in[I_1]\times[I_2]\times[I_3]} \widetilde{u}(\omega_1)\widetilde{v}(\omega_2\omega_3)\big[e_\omega \mathbb{1}(|e_\omega| \le \delta_0) - \mathbb{E} e_\omega \mathbb{1}(|e_\omega| \le \delta_0)\big],
\end{aligned} \tag{40}$$

which is a sum of independent centered random variables. Observe that

$$\left|\widetilde{u}(\omega_1)\widetilde{v}(\omega_2\omega_3)\big[e_\omega \mathbb{1}(|e_\omega|\le\delta_0)-\mathbb{E}e_\omega\mathbb{1}(|e_\omega|\le\delta_0)\big]\right|\le 2\delta_0|\widetilde{u}(\omega_1)||\widetilde{v}(\omega_2\omega_3)|,\quad \forall\omega$$

implying that each term in RHS of (40) is a bounded random variable. By applying Hoeffiding's inequality (**?**) to (40), we get

$$\begin{aligned}\mathbb{P}\Big(\Big|\widetilde{\mathbf{u}}^\top\big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty\le\delta_0)-\mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty\le\delta_0)\big]\widetilde{\mathbf{v}}\Big|\ge t\Big)\le& 2\exp\Big(\frac{-2t^2}{4\delta_0^2\sum_\omega\widetilde{u}(\omega_1)^2\widetilde{v}(\omega_2\omega_3)^2}\Big)\\ \le& 2\exp(-t^2/(2\delta_0^2)).\end{aligned}$$

Since the cardinality of the product set of $\mathfrak{D}_{1/3},\mathfrak{R}_{1/3}(J_1),\mathfrak{R}_{1/3}(R_2R_3)$ is bounded by $3^{C_0'(J_3R_3+J_1+R_2R_3)}$ for some absolute constant C_0', we apply a union bound and get

$$\begin{aligned}&\mathbb{P}\Big(\max_{\mathbf{D}\in\mathfrak{D}_{1/3}}\max_{\substack{\mathbf{u}\in\mathfrak{R}_{1/3}(J_1)\\ \mathbf{w}\in\mathfrak{R}_{1/3}(R_2R_3)}}\Big|\mathbf{u}^\top\mathbf{U}_1^\top\big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty\le\delta_0)-\mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty\le\delta_0)\big](\frac{\mathbf{G}_2}{\sqrt{I_2}}\otimes\mathbf{U}_3\mathbf{D})\mathbf{w}\Big|\ge t\Big)\\ &\quad\le 2\exp\left(-\frac{t^2}{2\delta_0^2}+C_1(J_3R_3+J_1+R_2R_3)\right)\overset{t=2\sqrt{C_1(J_1+J_3R_3+R_2R_3)}\delta_0\log^{1/2}I_1}{\le}2I_1^{-2}\end{aligned}$$

implying that with probability at least $1-2I_1^{-2}$,

$$\begin{aligned}&\max_{\mathbf{D}\in\mathfrak{D}_{1/3}}\big\|\mathbf{U}_1^\top\big[\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty\le\delta_0)-\mathbb{E}\mathcal{M}_1(\mathcal{E})\mathbb{1}(\|\mathcal{E}\|_\infty\le\delta_0)\big](\mathbf{G}_2\otimes\mathbf{U}_3\mathbf{D})\big\|/\sqrt{I_2}\\ &\quad\le C_1\sqrt{J_1+J_3R_3+R_2R_3}\log^{3/2}I_1.\end{aligned}\tag{41}$$

Putting together (41), (39), (38), (37) and (36), we conclude that with probability at least $1-4I_1^{-2}$,

$$\sup_{\mathbf{A}\in\mathfrak{B}(J_3,R_3)}\big\|\mathbf{U}_1^\top\mathcal{M}_1(\mathcal{E})(\mathbf{G}_2\otimes\mathbf{U}_3\mathbf{A})\big\|/\sqrt{I_2}\le C_3\sqrt{J_1+J_3R_3+R_2R_3}\log^{3/2}I_1$$

for some absolute constant $C_3>0$. This proves (11) of Lemma 4. □

Proof of Lemma 6. Recall by definition that $\widetilde{\mathbf{G}}_m^\top\widetilde{\mathbf{G}}/I_m=\mathbf{I}_{R_m}$. Moreover, the column space

of $\widetilde{\mathbf{G}}_m$ is a subspace of column space of $\mathbf{\Phi}_m(\mathbf{X}_m)$. Denote $\mathbf{U}_m \in \mathbb{R}^{I_m \times J_m}$ the left singular vectors of $\mathbf{\Phi}_m(\mathbf{X}_m)$. Then, there exists a $\widetilde{\mathbf{B}}_m \in \mathbb{R}^{J_m \times R_m}$ so that $\widetilde{\mathbf{G}}_m = \mathbf{U}_m \widetilde{\mathbf{B}}_m$ and $\widetilde{\mathbf{B}}_m^\top \widetilde{\mathbf{B}}_m / I_m = \mathbf{I}_{R_m}$ where $\widetilde{\mathbf{B}}_m$ is dependent with $\mathcal{E}$ while $\mathbf{U}_m$ is independent with $\mathcal{E}$. Denote $\mathcal{G}(d_1, d_2) := \{\mathbf{B} \in \mathbb{R}^{d_1 \times d_2}, \mathbf{B}^\top \mathbf{B} = \mathbf{I}_{d_2}\}$. Then,

$$\frac{\left\|\mathcal{M}_1\left(\mathcal{E} \times_1 \widetilde{\mathbf{G}}_1^\top \times_2 \widetilde{\mathbf{G}}_2^\top \times_3 \widetilde{\mathbf{G}}_3^\top\right)\right\|}{\sqrt{I_1 I_2 I_3}} \leq \sup_{\widetilde{\mathbf{B}}_m \in \mathcal{G}(J_m, R_m)} \left\|\mathcal{M}_1\left(\mathcal{E} \times_1 (\mathbf{U}_1 \widetilde{\mathbf{B}}_1)^\top \times_2 (\mathbf{U}_2 \widetilde{\mathbf{B}}_2)^\top \times_3 (\mathbf{U}_3 \widetilde{\mathbf{B}}_3)^\top\right)\right\|.$$

By choosing 1/5-nets of $\mathcal{G}(J_1, R_1), \mathcal{G}(J_2, R_2)$ and $\mathcal{G}(J_3, R_3)$ (e.g., by the covering number of Grassmannians in **?**), respectively, we can reproduce the proof of Lemma 4 and show that with probability at least $1 - I_1^{-2}$,

$$\left\|\mathcal{M}_1\left(\mathcal{E} \times_1 \widetilde{\mathbf{G}}_1^\top \times_2 \widetilde{\mathbf{G}}_2^\top \times_3 \widetilde{\mathbf{G}}_3^\top\right)\right\| \leq C_1 \sqrt{J_1 R_1 + J_2 R_2 + J_3 R_3 + R_1 + R_2 R_3} \log^{3/2} I_1$$

for some absolute constant $C_1 > 0$. Since $J_1 \asymp J_2 \asymp J_3$ and $R_1 \geq R_2 \geq R_3$, we can simplify the upper bound to $C_1' \sqrt{J_1 R_1} \log^{3/2} I_1$ and finish the proof of Lemma 6. □

Appendix F More real data applications

F.1 Human Brain Connection Data

As an additional illustration of using the STEFA and IP-SVD for explanatory data analysis, we consider partitioning the brain connectivity according to the 136×4 covariate-relevant loading matrix $\mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$. As mentioned in the main text, we choose $\mathbf{\Phi}_3(\mathbf{X})$ generated by polynomial basis of order 1, which is a similar linear setting as that in **?** with identity link function. Each column of $\mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$ corresponds to one of the four directions with largest variance among individual features in $\mathbf{X}$. The meaning of each latent direction can be inferred from $\mathbf{B}_3$. Figure 1 presents the heat map of $\mathbf{B}_3$, showing that the four factors weight mostly on the four columns of $\mathbf{\Phi}_3(\mathbf{X})$, namely, all-ones vector, female variable, and Age 22-25 variable and 31+ variable, respectively. Then each factor is interpreted as the

effects associated with global average, female, Age 22-25 variable and 31+.

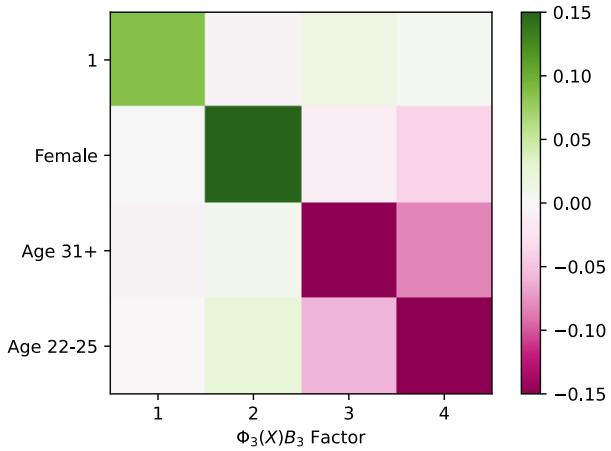
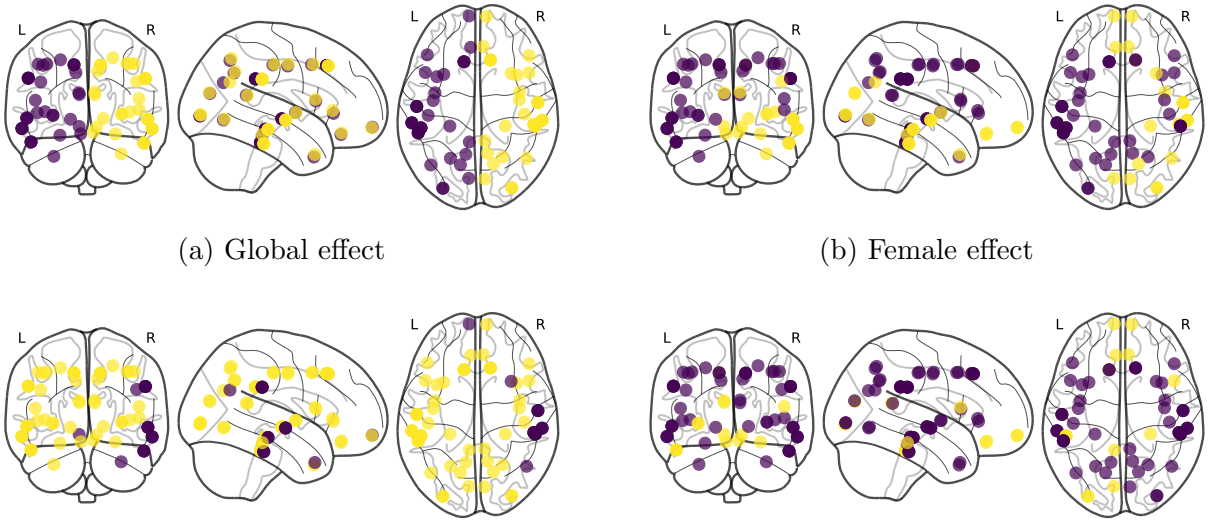


Figure 1: Heat map of the coefficient $\mathbf{B}_3$ in the covariate-relevant loading with polynomial basis of order one.

We obtained a $68 \times 68 \times 4$ connectivity tensor by projecting the original connectivity tensor on the column space of $\mathbf{\Phi}_3(\mathbf{X})\mathbf{B}_3$. As a result, the four slices along the third mode corresponds to the connectivity matrix for each of the global average, female, Age 22-25 variable and 31+ effects. We divide the 68 regions of the brain into two clusters based on the connectivity matrix for each of the effects. The clustering results are plotted in Figure 2. The connection within the same cluster is stronger than that between clusters. Some connectivity patterns can be observed. For example, the global connection exhibits clear left and right spatial separation and the age 22 - 25 group shows additional inter-hemispheric connectivity. While such explanatory analysis can provide some interesting observation, more rigorous methods, such as statistical testing procedures, need to be developed to support any scientific claim. These are important directions for future statistics researches.



(a) Global effect

(b) Female effect

(c) Age 22-25

(d) Age 31+

Figure 2: Partition of brain regions for each latent dimension corresponding to the covariate-relevant loading.