Guaranteed Functional Tensor Singular Value Decomposition

Rungang Han\textsuperscript{a}, Pixu Shi\textsuperscript{b}, and Anru R. Zhang\textsuperscript{c}

\textsuperscript{a}Department of Statistical Science, Duke University, Durham, NC; \textsuperscript{b}Department of Biostatistics & Bioinformatics, Duke University, Durham, NC; \textsuperscript{c}Department of Biostatistics & Bioinformatics, Computer Science, Mathematics, and Statistical Science, Duke University, Durham, NC

\textbf{ABSTRACT}

This article introduces the functional tensor singular value decomposition (FTSVD), a novel dimension reduction framework for tensors with one functional mode and several tabular modes. The problem is motivated by high-order longitudinal data analysis. Our model assumes the observed data to be a random realization of an approximate CP low-rank functional tensor measured on a discrete time grid. Incorporating tensor algebra and the theory of reproducing kernel Hilbert space (RKHS), we propose a novel RKHS-based constrained power iteration with spectral initialization. Our method can successfully estimate both singular vectors and functions of the low-rank structure in the observed data. With mild assumptions, we establish the non-asymptotic contractive error bounds for the proposed algorithm. The superiority of the proposed framework is demonstrated via extensive experiments on both simulated and real data. Supplementary materials for this article are available online.

\section{1. Introduction}

In recent decades, the analysis of tensor data has become an active research topic. Datasets in the form of high-order tensors or multiway arrays arise from various scientific applications, such as neuroimaging (Zhang et al. 2019), microscopy imaging (Zhang et al. 2020), and longitudinal microbiome study (Martinno et al. 2021). Such high-order data pose significant challenges in both theoretical analysis and computation implementation due to their complicated structures and a large number of entries involved, making it inappropriate to extend many existing methods for matrices to the analysis of tensor data.

In real applications, the different modes of tensor datasets (or different directions that the tensor arrays align in) can come in various formats. Two prominent formats are (a) \textit{tabular modes}, such as subject ID, genomics ID, treatments, where shuffling the indices does not essentially change the data structure; (b) \textit{functional modes}, such as time, location, spectrum (in the hyperspectral imaging), where the order of indices exhibits structures such as continuity. Here we provide two scenarios to illustrate tensor data with both types of modes.

\begin{itemize}
  \item \textbf{Multivariate functional data analysis}. Data with different functional features appear in various applications and can be formatted into a tensor with two tabular modes representing units and variables, and a functional mode representing time point. One example is the longitudinal microbiome studies, where microbiome samples are taken from multiple subjects (units) at multiple time points to study the abundance of bacteria (variables) over time. Depending on what taxonomic level is being studied, there can be hundreds and thousands of bacterial taxa in the feature mode and many of these taxa have strong correlation in their abundance.
  \item \textbf{Dynamic networks}. In network analysis, one often observes multiple snapshots of dynamic networks. The adjacency matrix in each snapshot can be stacked together into an adjacency tensor, where two tabular modes correspond to the vertices of the network, and a functional mode corresponds to time. The dimension of the tabular modes, that is, number of nodes, can be as large as thousands, but the hidden community structure of the network can often be modeled by low-rankness of adjacency tensor.
\end{itemize}

Just as in the examples above, the tabular modes may be very high-dimensional. The functional mode is often continuous, which is essentially infinite-dimensional. Therefore, dimension reduction tools are crucial for the visualization and analysis of high-dimensional tensor data with both tabular modes and functional modes. Various lines of methods can be applied, whereas one prominent framework is the \textit{multivariate/multilevel functional principal component analysis} (MFPCA) in the functional data analysis (FDA) literature. The classic MFPCA methods often focus on a moderately large number of variables and dimension reduction on the functional mode. The resulting number of estimated eigenfunctions is often in the same order as the number of variables, making the eigenfunctions hard to interpret when the variable mode is high-dimensional. In addition, most existing MFPCA frameworks focus on characterizing the covariance structure among functional variables at the population level and assuming the iid observations. Such assumption can be violated when dealing with heterogeneous...
samples. A more detailed discussion on the literature of multivariate functional analysis is given in Section 1.2. Another important class of works uses the high-order structure of the tensor data via different types of low-rank tensor decomposition models including (sparse) CP low-rankness (Anandkumar, Ge, and Janzamin 2014; Sun et al. 2017), (sparse) Tucker low-rankness (Zhang and Xia 2018; Zhang and Han 2019). While the low-rank models achieve efficient dimension reduction and are interpretable when the modes are tabular, they do not directly characterize or use the information included in the functional modes such as time and location. Therefore, new methods and theory are needed to tackle both the high-dimensional and functional aspects of functional tensor data.

In this article, we introduce a new framework for the dimension reduction of functional tensor data which we refer to as Functional Tensor Singular Value Decomposition (FTSVD). Suppose $Y \in \mathbb{R}^{p_1 \times p_2 \times T}$ is the underlying function of interest, where $T \subset \mathbb{R}$ is some compact set on the real axis. For any fixed $(i,j)$ pair, $Y_{ijt}$ is a function of $t \in T$. We aim to decompose $Y$ into the following canonical polyadic (CP) format:

$$
Y = X + Z,	ext{ where } X = \sum_{l=1}^{r} \lambda_l a_l \circ b_l \circ \xi_l \text{ or equivalently } 
X_{ijt} = \sum_{l=1}^{r} \lambda_l (a_l)_i (b_l)_j \xi_l(t).
$$

Here, $a_l \in \mathbb{R}^{p_1}$, $b_l \in \mathbb{R}^{p_2}$ are singular vectors of tabular modes; $\xi_l : T \rightarrow \mathbb{R}$ is the singular function, which corresponds to the eigenfunction in functional PCA; $Z$ is the remainder term of the rank-r CP decomposition of $Y$. We assume the observed data $\tilde{Y}$ to be $Y$ measured over a discrete grid $\{s_k\}_{k=1}^{n} \subset T$ with observational noise, that is,

$$
\tilde{Y} \in \mathbb{R}^{p_1 \times p_2 \times n}, \quad \tilde{Y}_{ijk} = Y_{ijk} + \varepsilon_{ijk}.
$$

Without loss of generality, we assume that $T = [0, 1]$ throughout the article.

The proposed model setting is flexible and adaptive. For example, in the scenario of multivariate functional data analysis, different from many existing literature on (multivariate) functional PCA (Happ and Greven 2018; Wang, Wong, and Zhang 2020), which assumed the samples are iid distributed and interpretability of subsequent analysis such as clustering, classification, and regression.

1.1. Contributions

The main contributions of this article are 3-fold. First, to the best of our knowledge, we are the first to study multivariate/multilevel functional data analysis through a lens of a low-rank tensor decomposition model. We are also the first to adopt the reproducing kernel Hilbert space (RKHS) framework for the singular function in the low-rank decomposition. Compared to recently emerged state-of-the-art tabular tensor models, the proposed model can better characterize the longitudinal behaviors of the data when there exist temporal modes. Compared to most classic FDA literature, our model does not assume the samples to be iid. Instead, our method may provide new insight on how to analyze heteroscedastic functional effects for mixed-population samples.

Second, we propose a new power iteration algorithm for the estimation of our model. In each iteration step, we use projection to update the estimators for the tabular components and constrained/regularized empirical risk minimization (ERM) to update the estimators for the functional component, respectively. The novel inclusion of multiple-times ERM in our iterative algorithm imposes new challenges on both statistical theory and numeric optimization.

Third, we establish the finite-sample statistical error bounds for the estimation of both singular vectors and functions (16). Our error bound includes contribution from the remainder term of CP approximation (Equation (1)) and observational noises (Equation (2)) relative to the signal level in $X$. It also reflects the effect of time grid density and incoherence condition of the CP decomposition. To the best of our knowledge, this is the first contractive error bound for iterative algorithms in functional data analysis.

1.2. Related Work

This work is related to a series of papers on low-rank tensor decomposition from statistics and machine learning, which aims at recovering the low-dimensional linear representations of the noisy tabular tensors (Anandkumar et al. 2014). Commonly-used low-rank models include CP (Anandkumar, Ge, and Janzamin 2014) and Tucker (Zhang and Xia 2018) where various power iteration methods were proposed on the singular vector (or principal component) estimation with statistical consistency guarantees. Motivated by real data applications in practice, related statistical models were further studied under non-Gaussian or missing-data scenarios. For example, Hong, Kolda, and Duersch (2020) and Han, Willett, and Zhang (2021) considered the likelihood-based generalized low-rank tensor decomposition under general exponential family.

In addition, various dimension-reduced structures are imposed on the singular vectors or subspace of the low-
rank tensor decomposition model to better capture other intrinsic properties of the data, such as sparsity (Sun et al. 2017; Zhang and Han 2019) and blocking (Han et al. 2020a). The most relevant paper to our work is the spatial/temporal structure, which is incorporated to characterize time/location-varying patterns for one or more of the tensor modes. There are various formulations for modeling such structures based on different real data motivations. For example, Sun and Li (2019) considered a fusion structure and assumed the time-varying trend to be piece-wise constant; Han et al. (2019) considered the multidimensional isotonic regression, where there are monotonous trends in each tensor mode. In most of such literature, the temporal structures are still assumed on tabular tensors and make the corresponding models restrictive. In contrast, the target of our framework is the underlying nonparametric functional structures, which reflects the trend of observed data over time.

Another related topic is the functional data analysis (FDA), which is a popular branch of statistical research. The readers are referred to books and surveys in Ramsay and Silverman (2006) and Wang, Chion, and Müller (2016). Starting from samples with univariate functions, traditional FDA involves various statistical learning tasks, such as regression (Yao, Müller, and Wang 2005; Cai and Yuan 2012), covariance function estimation (Rice and Silverman 1991), principal component analysis (PCA) (James, Hastie, and Sugar 2000), etc. Under the multivariate regime, Hasenstab et al. (2017) considered the multilevel/multidimensional functional data PCA. Wang, Wong, and Zhang (2020) studied the low-rank covariance estimation for multidimensional functional data. Fan, James, and Radchenko (2015) considered the functional additive regression method for high-dimensional functional regression. Hu and Yao (2021) considered the dimension reduction via dynamic principal subspace for multivariate functional data. Chen, Delicado Useros, and Müller (2017) studied the functional data analysis for scenarios where the observations are functional at each location. The works on multivariate functional principal component analysis (Allen 2013; Happ and Greven 2018) also address dimension reduction of multidimensional functional data. Among these literature, Happ and Greven (2018) and Wang, Wong, and Zhang (2020) are the most relevant to our work. In particular, Wang, Wong, and Zhang (2020) focused on iid p-dimensional function objects, and they imposed tensor low-rankness on the corresponding (2p)-dimensional covariance functions; Happ and Greven (2018) considered the multivariate functional PCA for iid samples from some multivariate functional population and the data can be formulated as a typical functional tensor stack by “sample-feature-function.” Both papers assumed iid samples and focused on the statistical inference in the covariance function. Compared to these methods, our model does not assume independence of the signal tensor along any mode. Such flexible modeling is able to encompass a range of problems on heterogeneous or heteroscedastic data. The resulting estimated singular vectors can also be used to perform many important tasks, such as clustering, classification, and regression.

Our framework is also related to a line of recent papers on high-order tensor factor analysis. For example, Chen, Yang, and Zhang (2021) and Han et al. (2020b) considered a factor model for dynamic tensor time series, where they assumed the factor tensors are (weakly) stationary without imposing any deterministic or structured time-varying trends. Chen et al. (2020) proposed a semiparametric Tucker decomposition model, where the loadings on one or more modes can be approximated by smooth functions from an H"older class. Our framework is in the spirit of CP decomposition, which allows for more general functional classes via RKHS theory.

2. Low-Rank Functional Tensor Decomposition Model

2.1. Notation and Preliminaries

We use the lowercase letters, for example, $a,b,u,v$, to denote scalars or vectors. For any $a,b \in \mathbb{R}$, let $a \land b$ and $a \lor b$ be the minimum and maximum of $a$ and $b$, respectively. For a vector $u \in \mathbb{R}^n$, $\|u\|_2$ denotes its Euclidean norm. The unit sphere in $\mathbb{R}^n$ is denoted as $S^{n-1}$. Matrices are denoted as uppercase letters such as $A$ and $B$. In addition, we let $\mathcal{O}_{p,r}$ be the collection of all $p$-by-$r$ matrices with orthonormal columns: $\mathcal{O}_{p,r} := \{U \in \mathbb{R}^{p \times r} : U^T U = I\}$, where $I$ is the identity matrix. We also denote $\mathcal{O}_{r} := \mathcal{O}_{r,r}$ as the set of all $r$-dimensional orthogonal matrices.

Let $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq 0$ be the singular values of $A$ in descending order and let $\text{SVD}(A)$ be the matrix comprised of the top $r$ left singular vectors of $A$. Let $\|A\|_F = \lambda_1(A)$ and $\|A\|_{F_2} = \sqrt{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} A_{ij}^2} = \sqrt{\sum_{i=1}^{p_1} \lambda_i(A)}$ be the matrix spectral and Frobenius norms, respectively. For any matrix $A = [a_1, \ldots, a_l] \in \mathbb{R}^{l \times 1}$ and $B \in \mathbb{R}^{k \times 1}$, the Kronecker product is defined as the $(IK)$-by-$(JL)$ matrix $A \otimes B = [a_1 \otimes B \cdots \otimes B]$. Without loss of generality, we assume the function domain of interest is $T = [0,1]$ and denote $L^2([0,1])$ as the functional space of all square-integrable functions on $[0,1]$, that is,

$$L^2([0,1]) = \{ f : [0,1] \to \mathbb{R} : \|f\|_{L^2} < \infty \},$$

where $\|f\|_{L^2} := \left( \int_0^1 f^2(t) dt \right)^{1/2}$.

Denote the inner product of any two functions $f,g \in L^2([0,1])$ as $(f,g)_{L^2} = \int_0^1 f(t)g(t) dt$. For a sequence $\{s_k\}_{k=1}^n$ in $[0,1]$, we denote $f_n := (f(s_1), \ldots, f(s_n)) \in \mathbb{R}^n$ and $\|f_n\|_n := \sqrt{\sum_{k=1}^{n} f^2(s_k)/n}$. We use $C, C_0, C_1, \ldots$ and $c, c_0, c_1, \ldots$ to represent generic large and small positive constants, respectively. The actual values of these generic symbols may differ from line to line. We denote $a \lesssim b$ if $a \leq Cb$ for a constant $C > 0$ that does not depend on other model parameters; we say $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$ both hold.

Throughout this article, the tensors are denoted by uppercase calligraphy letters, such as $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. For a tabular tensor $A \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_l}$, let $A_{ijk}$ be the $(i,j,k)$th entry. For any $u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}, w \in \mathbb{R}^{p_3}$, the mode-1 tensor-vector product is defined as $A \times_1 u \in \mathbb{R}^{p_2 \times p_3}$, $(A \times_1 u)_{jk} := \sum_{i=1}^{p_1} u_i A_{ijk}$. The mode-2 and mode-3 tensor-vector products, $A \times_2 v$ and $A \times_3 w$ for $v \in \mathbb{R}^{p_2}$ and $w \in \mathbb{R}^{p_2}$, can be defined in the parallel way. The multiplication along different modes is cumulative and commutative. For example, $A \times_1 u \times_2 v \in \mathbb{R}^{p_3}$ with $(A \times_1 u \times_2 v)_{k} = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} u_i v_j A_{ijk}$; and $A \times_1 u \times_2 v \times_3 w = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} u_i v_i w_k A_{ijk} \in \mathbb{R}$. We also introduce the matricization operator that transforms tensors to matrices.
Particularly, the mode-1 matricization of $A \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is defined as $M_1(A) \in \mathbb{R}^{p_1 \times (p_2 p_3)}$, where $[M_1(X)]_{ij} = \tilde{A}_{ijk}$. In other words, each row of $M_k(A)$ is the vectorization of a mode-$k$ slice.

Next, we extend the notions of tabular tensors to tensors with a hybrid of tabular and functional modes. We still denote the tensors with hybrid modes by uppercase calligraphy letters. Suppose $A = [A_{ijr}, i \in [p_1], j \in [p_2], r \in [0,1]] \in \mathbb{R}^{p_1 \times p_2 \times [0,1]}$ is an order-3 tensor with two tabular modes and a functional mode. The tensor-vector products on the tabular modes are defined as convention: $A \times_1 u \in \mathbb{R}^{p_1 \times [0,1]}$, $(A \times_1 u)b := \sum_{i=1}^{p_1} u_i A_{ijr}$. Suppose $f \in L^2([0,1])$, we define the tensor-functional product as $A \times_3 f \in \mathbb{R}^{p_1 \times p_2}$, $(A \times_3 f)(j) := \int_0^1 A_{ijr}f(t)dt$. The multiple-mode tensor-vector/function multiplication can be similarly defined by combining the operators of single-mode multiplication. For example, $A \times_1 u \times_2 v$ yields a function $\sum_{j=1}^{p_2} \sum_{i=1}^{p_1} u_i v_j A_{ijr}$ and $A \times_1 a \times_2 b \times_3 f$ yields a scalar $\sum_{j=1}^{p_2} \sum_{i=1}^{p_1} \int_0^1 u_i v_j A_{ijr}f(t)dt$.

Next, we provide the preliminaries for the reproducing kernel Hilbert space (RKHS). Consider a Hilbert space $\mathcal{H} \subset L^2([0,1])$ with the associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We assume there exists a continuous symmetric positive-semidefinite kernel function $K : [0,1] \times [0,1] \rightarrow \mathbb{R}^+$ that satisfies the following RKHS conditions: (1) for any $s \in [0,1]$, $K(\cdot, s) \in \mathcal{H}$; (2) for each $g \in \mathcal{H}$, $g(t) = \langle g, K(\cdot, t) \rangle_{\mathcal{H}}$, $\forall t \in [0,1]$. By Mercer’s Theorem (Mercer 1909), there exists an orthonormal basis $\{\phi_k\}_{k=1}^{\infty}$ of $L^2([0,1])$, such that $K(\cdot, s)$ admits the following eigenfunction decomposition: $K(s,t) = \sum_{k=1}^{\infty} \mu_k \phi_k(s)\phi_k(t)$, $\forall s, t \in [0,1]$. Here, $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are the nonnegative eigenvalues of $K$. We assume $K$ has a finite trace norm: $\sum_{k=1}^{\infty} \mu_k < \infty$. Then any $f \in \mathcal{H}$ can be decomposed on $\{\phi_k\}_{k=1}^{\infty}$ as

$$f(t) = \sum_{k=1}^{\infty} a_k \phi_k(t), \quad a_k := \int_0^1 f(s)\phi_k(s)ds, \quad t \in [0,1].$$

The RKHS norm of $f$ can be then represented as $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}} = \sqrt{\sum_{k=1}^{\infty} a_k^2 / \mu_k}$. Accordingly, for any two functions $f = \sum_{k=1}^{\infty} a_k \phi_k$ and $g = \sum_{k=1}^{\infty} b_k \phi_k$ in $\mathcal{H}$, the inner product with respect to $\mathcal{H}$ can be calculated as $\langle f, g \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} a_k b_k / \mu_k$.

Assume the following regularity condition on $K$ throughout the article.

**Assumption 1.** (1) $K$ satisfies $\max\{\mu_1, \sup_{s \in [0,1]} K(\cdot, s)\} \leq 1$; (2) there exists an absolute constant $C_K$, which only depends on $\mathcal{H}$, such that

$$\|g\|_{\mathcal{H}} \leq C_K \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}, \quad \forall f, g \in \mathcal{H}. \quad (3)$$

Assumption 1(a) can be ensured by rescaling $K$, that is, multiplying $K$ by some positive constant. This condition also implies

$$\|f\|_{L^2} \leq \|f\|_{\infty} := \sup_{t \in [0,1]} |f(t)| = \sup_{s \in [0,1]} |\langle f(\cdot), K(\cdot, s) \rangle_{\mathcal{H}}| \leq \|f\|_{\sup} \sup_{s \in [0,1]} \|K(\cdot, s)\|_{\mathcal{H}} = \|f\|_{\sup} \sup_{s \in [0,1]} \langle K(\cdot, s), (K(\cdot, s), s) \rangle_{\mathcal{H}}^{1/2} = \|f\|_{\sup} \sup_{s \in [0,1]} \sqrt{\langle K(s, s) \rangle_{\mathcal{H}}}. \quad (4)$$

Assumption 1(b) is introduced for convenience of the later theoretical analysis. This condition appears in the literature on high-dimensional functional linear regression/auto-regression (Wang et al. 2020a, 2020b) and also holds for some prevalent functional spaces of interest. For example, Wang et al. (2020a, 2020b) showed that the Sobolev space

$$W^{r,2} = \left\{ f : f^{(r)} \right\}_{r=0}^\infty \in L^2((0,1)) \right\}$$

satisfies (3) with some constant $C_K$ that only depends on $\alpha$ and showed in particular that $W^{1,2}$ satisfies (3) with $C_K = \sqrt{\pi}$. Define the effective dimension of $\mathcal{H}$ as $p_{\mathcal{H}} := \sum_{k=1}^{\infty} \mu_k^2 / \mu_1^2$. Intuitively speaking, $p_{\mathcal{H}}$ captures the intrinsic dimension of the Hilbert space $\mathcal{H}$ in the sense that the Gaussian processes over $\mathcal{H}$ has the similar incoherence as the Gaussian random vectors in a $p_{\mathcal{H}}$-dimensional Euclidean space (see the forthcoming Proposition 2 for details). As a concrete example, suppose $\mu_k = 1$ for $k \leq d$ and $\mu_k = 0$ for $k > d$. Then, $p_{\mathcal{H}} = d$ and $\mathcal{H}$ is equivalent to a $d$-dimensional Euclidean space.

### 2.2. Functional Tensor Decomposition

Let $Y \in \mathbb{R}^{p_1 \times p_2 \times [0,1]}$ be the functional tensor of interest. Assume $Y$ is approximately CP rank-$r$.

$$Y = X + Z, \quad \mathcal{X} = \sum_{l=1}^{r} \lambda_l a_l \otimes b_l \otimes \xi_l \in \mathbb{R}^{p_1 \times p_2 \times [0,1]}.$$

Here, each tuple $(\lambda_l, a_l, b_l, \xi_l)$ corresponds to a singular component of $\mathcal{X}$: $\lambda_l$ is the singular value, $a_l, b_l$ are the singular vectors, and $\xi_l$ is the singular function. We also assume $\|a_l\|_2 = \|b_l\|_2 = \|\xi_l\|_C = 1$ and $\lambda_l > 0$ for scaling identifiability. We denote $\lambda_{\max} = \max_{l \leq r} \lambda_l$ and $\lambda_{\min} = \min_{l \leq r} \lambda_l$. The functional tensor $Z \in \mathbb{R}^{p_1 \times p_2 \times [0,1]}$ is included to model the unexplained remainder term. We also assume $\xi_l$ satisfies the regularity condition, such that $\|\xi_l\|_{\mathcal{H}} \leq C_3$ for some given RKHS $\mathcal{H}$ and some universal constant $C_3$.

Motivated by the previously discussed applications, we assume the functional tensor is measured on a discrete grid of time points $\{s_k\}_{k=1}^{n_0}$ approximately distributed $\mathbb{R}^{p_1 \times p_2 \times n}$ is observed:

$$\tilde{Y}_{ijk} := Y_{ijk} + e_{ijk}, \quad i \in [p_1], j \in [p_2], k \in [n]. \quad (6)$$

Here, $e_{ijk}$ are independent Gaussian random noises with mean zero and variance $\tau^2$. Our goal is to perform functional tensor singular value decomposition on $\tilde{Y}$, that is to estimate $(\{a_l, b_l, \xi_l\})_{l=1}^{r}$ based on $\tilde{Y}$ and $(s_k)$, from the model (5) and (6).

### 2.3. Model Identifiability

We first discuss the identifiability conditions for the FTSVD model. We say the parameter tuple $(\lambda_l, a_l, b_l, \xi_l)_{l=1}^{r}$ is identifiable if and only if for any other parameter tuple $(\tilde{\lambda}_l, \tilde{a}_l, \tilde{b}_l, \tilde{\xi}_l)_{l=1}^{r}$ satisfying $\sum_{l=1}^{r} \lambda_l a_l \otimes b_l \otimes \xi_l = \sum_{l=1}^{r} \tilde{\lambda}_l \tilde{a}_l \otimes \tilde{b}_l \otimes \tilde{\xi}_l$ there exists some permutation $\pi$ on $[r]$ such that

$\tilde{\lambda}_l = \tilde{\lambda}_{\pi(l)}, \quad \tilde{a}_l \otimes \tilde{b}_l \otimes \tilde{\xi}_l = \tilde{a}_{\pi(l)} \otimes \tilde{b}_{\pi(l)} \otimes \tilde{\xi}_{\pi(l)}, \quad \forall l \in [r].$
In other words, all rank-1 components can be uniquely determined up to permutation and all factors are identifiable up to sign-flipping and permutation. The following Proposition 1 shows almost all parameter tuples are identifiable for moderate tensor rank $r$. Proposition 1 is in parallel with the identifiability condition of tabular tensor CP decomposition (Kruskal 1976).

**Proposition 1.** The set of unidentifiable parameter tuples has measure zero with respect to the Lebesgue measure on the parameter space $\Theta = \{(l_i, a_i, b_i, c_i)_{i=1}^r : l_i > 0, \|a_i\|_2 = \|b_i\|_2 = \|\xi_i\|_{L^2}, \xi_i \in \mathcal{H}\}$ when either of the following conditions is met:

- $\mathcal{H}$ is a functional space with finite dimension $p_3$, and $2r < p_1 + p_2 + p_3 - 2$;
- $\mathcal{H}$ is an infinitely dimensional space, all the $\xi_i$s are continuous, and $r < p_1 + p_2 - 2$.

### 3. Methods for Functional Tensor SVD

**Power iterations.** As a starting point, we approach the problem from an optimization perspective:

\[
\min_{\lambda, a_i, b_i, \xi_i} \sum_{i,j,k} \left( \hat{Y}_{ijk} - \sum_{r=1}^{\tau} \lambda_r \cdot (a_r)_{ij} \cdot (b_r)_{jk} \cdot \xi_r(s_k) \right)^2,
\]

subject to $\|a_i\|_2 = \|b_i\|_2 = \|\xi_i\|_{L^2} = 1, \|\xi_i\|_{\mathcal{H}} \leq C_{\xi_i}.$

However, the computation of (7) is highly nontrivial and the exact solution is computationally intractable in general since (a) the problem is highly non-convex or even NP-hard due to the tensor product and the multilinear structure; (b) (7) is essentially an infinite-dimensional optimization problem due to the functional argument $\xi_i$. Therefore, we introduce the following RKHS-based constrained power iteration method to overcome these difficulties. As a popular and powerful method for singular value decomposition, power iteration has been successfully applied to tabular tensor decomposition in the past two decades (Anandkumar et al. 2014; Anandkumar, Ge, and Janzamin 2014).

We first discuss the power iteration for the one-component case (i.e., $r = 1$), while the multiple-component scenario is postponed later. Given estimates at Step $t$: $(a^{(t)}, b^{(t)}, \xi^{(t)})$, power iteration updates $a^{(t)}$, $b^{(t)}$, and $\xi^{(t)}$ alternatively. Specifically, recall $\hat{Y}_n = \{\xi^{(t)}(s_1), \ldots, \xi^{(t)}(s_n)\}$ is the discretization of the function $\xi^{(t)}$. We update the singular vectors $a^{(t)}, b^{(t)}$ by projection-normalization:

\[
\hat{a}^{(t+1)} = \hat{Y} \times_{1} b^{(t)} \times_{3} \xi_n^{(t)}, \quad a^{(t+1)} = \hat{a}^{(t+1)}/\|\hat{a}^{(t+1)}\|_2, \\
\hat{b}^{(t+1)} = Y \times_{1} a^{(t)} \times_{3} \xi_n^{(t)}, \quad b^{(t+1)} = \hat{b}^{(t+1)}/\|\hat{b}^{(t+1)}\|_2. 
\]

Since only a discrete subset of observations from the functional tensor are accessible, we consider the following optimization to update the singular function estimator:

\[
\hat{\xi}^{(t+1)} = \arg \min_{\|\xi\|_{\mathcal{H}} \leq (C_\xi \lambda_{\max})} \sum_{i,j,k} \left( \hat{Y}_{ijk} - \left(a^{(t)}\right)_{ij} \cdot \left(b^{(t)}\right)_{jk} \cdot \xi(s_k) \right)^2, \\
\hat{\xi}^{(t+1)} = \hat{\xi}^{(t+1)}/\|\hat{\xi}^{(t+1)}\|_{L^2}. 
\]

Note that the first part of (9) is essentially a weighted mean functional estimation problem, or a special case of functional linear regression. Since $\hat{\xi}^{(t+1)}$ is essentially a minimizer of a regularized empirical risk function defined over the RKHS $\mathcal{H}$, it admits a finite-dimensional closed-form solution by the classic Representer Theorem (Kimeldorf and Wahba 1971) to be discussed in Section 5.

Let $(\hat{a}, \hat{b}, \hat{\xi})$ be the final estimates after sufficient iterations. The singular value is estimated via $\hat{\lambda} = \hat{Y} \times_1 \hat{a} \times_2 b \times_3 \hat{\xi}_n$. The procedure of one-component power iteration is summarized to Algorithm 1.

**Algorithm 1 Regularized Power Iteration**

**Require:** Tensor $\hat{Y} \in \mathbb{R}^{p_1 \times p_2 \times n}$, initialization $(a^{(0)}, b^{(0)})$, iteration number $T$

Calculate $\hat{\xi}_0$ via:

\[
\hat{\xi}_0 = \arg \min_{\|\xi\|_{\mathcal{H}} \leq (C_\xi \lambda_{\max})} \sum_{i,j,k} \left( \hat{Y}_{ijk} - (a^{(0)})_{ij} \cdot (b^{(0)})_{jk} \cdot \xi(s_k) \right)^2,
\]

\[
\hat{\xi}^{(0)} = \hat{\xi}_0 / \|\hat{\xi}_0\|_{L^2}.
\]

**for all** $l = 0, T - 1$

Calculate $a^{(l+1)}, b^{(l+1)}$ via (8) and calculate $\xi^{(l+1)}$ via (9).

**end for**

Calculate $\hat{\lambda}^{(T)} = \hat{Y} \times_1 a^{(T)} \times_2 b^{(T)} \times_3 \hat{\xi}_n^{(T)}$, $\hat{\lambda} = \hat{\lambda}^{(T)}$.

**return** $(\hat{\lambda}, \hat{a}, \hat{b}, \hat{\xi}) = (\hat{\lambda}^{(T)}, a^{(T)}, b^{(T)}, \xi^{(T)})$.

**Initialization and Overall Algorithm.** Next, we discuss initialization scheme for $(a^{(0)}, b^{(0)})$, which is required for the implementation of Algorithm 1. If $\mathcal{X}$ is a rank-1 tensor, the matricizations of the signal tensor $\mathcal{X}$ admit the following rank-1 decompositions:

\[
\mathcal{M}_1(\mathcal{X}) = \lambda_1 a_1 \otimes (\xi_1)_{n}, \quad \mathcal{M}_2(\mathcal{X}) = \lambda_1 b_1 \otimes (\xi_1)_{n}^\top, 
\]

that is, $a_1$ and $b_1$ are the left singular vectors of $\mathcal{M}_1(\mathcal{X})$ and $\mathcal{M}_2(\mathcal{X})$, respectively. Recall “$\otimes$” is the Kronecker product and $b_1 \otimes (\xi_1)_{n} \in (\mathcal{R}^{p_2})^{\times n}$-dimensional vector. Given $\hat{Y}$ being a noisy substitute of $\mathcal{X}$, it is natural to initialize $a_1$ and $b_1$ by the first singular vectors of $\mathcal{M}_1(\hat{Y})$ and $\mathcal{M}_2(\hat{Y})$, respectively:

\[
a^{(0)} = \text{SVD}_1(\mathcal{M}_1(\hat{Y})), \quad b^{(0)} = \text{SVD}_1(\mathcal{M}_2(\hat{Y})).
\]

Our estimation procedure for the rank one scenario is then completed by combining Algorithm 1 with Equation (10). When the rank $r > 1$, there are multiple singular components to be estimated and the power iteration will no longer be the exact alternating minimization scheme. Nevertheless, when all the singular components $((a_i, b_i, \xi_i))_{i=1}^r$ satisfy the incoherent condition and $(a^{(0)}, b^{(0)})$ satisfies some initialization conditions, Algorithm 1 still yields estimators with guaranteed local convergence. Therefore, we propose to perform spectral initialization scheme sequentially: we first let $a^{(0)}$ and $b^{(0)}$ be initialized via (10) and let $(\hat{\lambda}_1, \hat{a}_1, \hat{b}_1, \hat{\xi}_1)$ be the first estimated singular component by power iteration (Algorithm 1). Then we subtract $\lambda_1 \hat{a}_1 \otimes \hat{b}_1 \otimes (\hat{\xi}_1)_n$ from the original observation $\hat{Y}$. Note that this is essentially equivalent to projecting $\hat{Y}$ onto the low-dimensional tensor space spanned by $\hat{a}_1 \otimes \hat{b}_1 \otimes (\hat{\xi}_1)_n$ and replacing the data with the projection residue. We perform this procedure...
for \( r \) times to obtain \( r \) estimated singular components. The pseudocode of this procedure is summarized to Algorithm 2.

**Algorithm 2** Functional Tensor SVD with Sequential Spectral Initialization

Require: Tensor \( \hat{\mathcal{Y}} \in \mathbb{R}^{p_1 \times p_2 \times n} \), rank \( r \)

for \( l = 1, \ldots, r \) do

Calculate \( a^{(l)} = \text{SVD}_l \left( \mathcal{M}_1(\hat{\mathcal{Y}}) \right), b^{(l)} = \text{SVD}_l \left( \mathcal{M}_2(\hat{\mathcal{Y}}) \right) \).

Calculate \( (\hat{\lambda}_l, \hat{a}_l, \hat{b}_l, \hat{\xi}_l) \) by applying Algorithm 1 on \( (a^{(l)}, b^{(l)}) \).

Update \( \hat{\mathcal{Y}} = \hat{\mathcal{Y}} - \hat{\lambda}_l \hat{a}_l \circ b_l \circ (\hat{\xi}_l)_n \).

end for

return \( \{(\hat{\lambda}_l, \hat{a}_l, \hat{b}_l, \hat{\xi}_l)\}_{l=1}^r \)

### 4. Statistical Theory

In this section, we present the theoretical results for the proposed estimation algorithms. We particularly aim to study the statistical error bound of \( \hat{a}, \hat{b}, \hat{\xi} \). Since the singular vectors and functions are identifiable up to sign flips, we focus on the following *sine values* of the pairs of vectors/functions:

\[
\text{dist}(u, v) = \sqrt{1 - \left( \frac{u^T v}{\|u\|_2\|v\|_2} \right)^2}, \quad \forall u, v \in \mathbb{R}^p;
\]

\[
\text{dist}(f, g) = \sqrt{1 - \left( \frac{\int_0^1 f(t)g(t)dt}{\|f\|_2\|g\|_2} \right)^2}, \quad \forall f, g \in \mathbb{L}^2([0, 1]).
\]

We introduce the following quantity associated with the RKHS \( \mathcal{H} \):

\[
\zeta_n := \inf \left\{ \xi \geq \sqrt{\log n/n} : Q_n(\delta) \leq \xi \delta + \xi^2, \quad \forall \delta \in (0, 1] \right\},
\]

where

\[
Q_n(\delta) := \left( \sum_{k=1}^{\infty} \min(\delta^2, \mu_k) \right)^{1/2}.
\]

and \( \mu_1 \geq \mu_2 \geq \cdots \geq 0 \) are the eigenvalues of the reproducing kernel \( \mathcal{K} \) of \( \mathcal{H} \). Essentially, \( \zeta_n \) quantifies the information loss from only observing measurements on a discrete grid, rather than the whole function, of \( f \) in the RKHS \( \mathcal{H} \). This quantity and its variants are commonly used in the literature on functional data analysis (Koltchinskii and Yuan 2010; Raskutti, Wainwright, and Yu 2012; Wang et al. 2020b). As the grid density \( n \) increases, \( \log n/l^2 \) reveals more information on \( f \) and the value of \( \zeta_n \) decreases. For some specific RKHS \( \mathcal{H} \), the explicit rate of \( \zeta_n \) can be obtained. As an example, when \( \mathcal{H} \) is the Hilbert–Sobolev space \( W^{\alpha,2}(0, 1) \) defined in (4), we have \( \zeta_n \lesssim n^{-\alpha/(2\alpha+1)} \) (see Proposition 6, supplementary materials). More technical results on \( Q_n(\delta) \) and \( \zeta_n \) will be discussed in Appendix C, supplementary materials.

We also introduce the following quantity to measure the scale of the unexplained remainder term \( \mathcal{Z} \):

\[
\mathcal{E} := \sup_{x \in [0, 1]} \| \mathcal{Z} x \|.
\]

Here, \( \| \cdot \| \) is the largest singular value of the matrix. \( \mathcal{E} \) can be interpreted as an extension of the *tensor spectral norm* from the tabular tensors, which appears as the prominent term in the estimation error bound in various statistical tensor models (Anandkumar, Ge, and Janzamin 2014; Sun et al. 2017; Han, Willett, and Zhang 2021). In particular, the forthcoming Proposition 3 explicitly characterizes \( \mathcal{E} \) for Gaussian processes \( \mathcal{Z}_j \) via the dimensions of the tabular modes \( p_1, p_2 \).

### 4.1. Assumptions

In this section, we present the technical assumptions that are used to establish the theoretical guarantees for the proposed algorithm. We first introduce the following incoherence conditions on the true singular components \( (a_j, b_j, \xi_j) \) of the signal tensor \( \mathcal{X} \).

**Assumption 2 (Incoherence).** Consider model (5). Assume

\[
\mu := \max \left\{ \max_{i \neq j} \| (a_i, a_j) \|, \max_{i \neq j} \| (b_i, b_j) \|, \max_{i \neq j} \| (\xi_i, \xi_j) \| \right\}
\]

\[
\leq c/(\kappa^2 r^2),
\]

where \( \kappa := \lambda_{\max}/\lambda_{\min} \) is the condition number.

**Assumption 2** suggest the singular components of \( \mathcal{X} \) are pairwise incoherent. Such the condition is widely assumed in the literature on tabular CP low-rank tensor decomposition. By (Anandkumar, Ge, and Janzamin 2014, Lemma 2), one can show that \( \max_{i \neq j} \| (a_i, a_j) \| \approx 1/\sqrt{p_1} \) and \( \max_{i \neq j} \| (b_i, b_j) \| \approx 1/\sqrt{p_2} \) with high probability if \( (a_i^T)_{i=1}^{p_1} \) and \( (b_i^T)_{i=1}^{p_2} \) are iid uniformly sampled from \( S^{p_1-1} \) and \( S^{p_2-1} \), respectively. We can also establish the incoherence for singular functions under certain sampling distributions described as follows.

**Proposition 2.** Suppose \( r \geq 2 \). Let \( f_i : \mathcal{G}(0, \cdot) \to \mathcal{G}(0, \cdot), i = 1, \ldots, r \) be iid mean-zero Gaussian processes with covariance function \( \mathcal{G}(s, t) = \sum_{k=1}^{\infty} \mu_k^2 \phi_k(s)\phi_k(t) \). Assume \( \log r < c \sum_{k=1}^{\infty} \mu_k^4/\mu_k^4 \) and normalize \( \xi_i = f_i/\|f_i\|_{\mathcal{L}^2} \). Recall the effective dimension \( p_\mathcal{H} := \sum_{k=1}^{\infty} \mu_k^2/\mu_k^4 \). Then, with probability at least \( 1 - Cr^{-8} \),

\[
\xi_i \in \mathcal{H}, \quad \| (\xi_i, \xi_j) \| \leq C \log r/\sqrt{p_\mathcal{H}}, \quad \forall 1 \leq i \neq j \leq r.
\]

Our next assumption is on the grid density and signal-to-noise ratio (SNR).

**Assumption 3 (Grid density and SNR).** The grid density \( n \) and the least singular value \( \lambda_{\min} \) satisfy

\[
\frac{n}{\log^2 n} \geq C (p_1 + p_2), \quad \zeta_n \leq c/\kappa^2 r,
\]

\[
\lambda_{\min} \geq C r \left( \mathcal{E} + r \left( \zeta_n + \sqrt{(p_1 + p_2)/n} \right) \right)
\]

(15)

for some sufficiently small constant \( c > 0 \) and large constant \( C > 0 \).

Recall that \( \zeta_n \) characterizes the difficulty of estimating a function in \( \mathcal{H} \) under discretization and Gaussian observational errors. In most interesting scenarios (which will be discussed later), these quantities have polynomial decay with respect to \( n \), and the first two conditions in (15) is mild that allows \( p, r \) to grow with \( n \). The third condition in (15) is similar to the
widely-assumed SNR condition in the literature on tabular tensor decomposition (Anandkumar, Ge, and Janzamin 2014; Han, Willett, and Zhang 2021), where the ratios of noise (e.g., \( \mathcal{E}, \tau \)) and the least singular value (\( \lambda_{\min} \)) are required to be sufficiently large. This guarantees that each singular component has strong-enough signal and can be estimated consistently.

**Assumption 4 (Initialization).** There exists some \( l \in [r] \) and constant \( c_0 > 0 \) such that the initialization error satisfies
\[
\max \{ \text{dist}(\hat{a}^{(0)}, a), \text{dist}(\hat{b}^{(0)}, b) \} \leq c_0/(kr).
\]

Assumption 4 requires that the initial tabular mode estimations are reasonably close to one of the true singular component. Such the condition is widely used in the theoretical analysis for power iteration (Anandkumar, Ge, and Janzamin 2014; Cai et al. 2019). In particular, when \( \mathcal{X} \) is a well-conditioned tensor with a constant rank (i.e., \( r = O(1) \)), this assumption means the initialization error is only smaller than a constant. Later, we will show the initial estimator yielded by the spectral initialization (Equation (10)) satisfies this assumption.

### 4.2. Local Convergence

The following theorem gives an estimation error bound for the proposed RKHS-based constrained power iteration (Algorithm 1).

**Theorem 1.** Suppose Assumptions 1–4 hold. Let \( a^{(t)}, b^{(t)}, \xi^{(t)} \) be estimated singular vectors and function at step \( t \) of Algorithm 1. Recall \( \mathcal{E} \) is defined in (14), \( \eta_t \) is defined in (12), \( \mu \) is defined in Assumption (2). Then with probability at least \( 1 - 2n^{-9} - C \log n \cdot e^{-c(p_1+p_2)} \), we have
\[
\max \{ \text{dist}(a^{(t)}, a), \text{dist}(b^{(t)}, b), \text{dist}(\xi^{(t)}, \xi) \} \\
\leq 2^{-t} + C \left( \frac{\mathcal{E}}{\lambda_{\min}} + \frac{\tau(\eta_t + (p_1 + p_2)/n)}{\lambda_{\min}} + \eta_t + (r - 1)\mu \right), \\
\forall t \geq 0. \tag{16}
\]

Specifically if \( r = 1 \), the estimation error of \( a^{(t)}, b^{(t)} \) enjoys the following better rate:
\[
\max \{ \text{dist}(a^{(t)}, a), \text{dist}(b^{(t)}, b) \} \\
\leq C \left( \frac{\mathcal{E}}{\lambda_{\min}} + \frac{\tau(\eta_t + (p_1 + p_2)/n)}{\lambda_{\min}} \right), \quad \forall t \geq 1.
\]

**Remark 1 (Interpretation of Estimation Error Bound).** Theorem 1 shows if the number of iterations \( T > \Omega(\log(1/\epsilon_{\text{fin}})) \), the estimation error of \( \hat{\xi} \) can be upper bounded by the sum of the following four terms: \( \mathcal{E}/\lambda_{\min}, \frac{\tau(\eta_t + (p_1 + p_2)/n)}{\lambda_{\min}}, \eta_t, \frac{\tau}{\lambda_{\min}} \cdot \left( \frac{(p_1p_2/n)^{1/4} + \sqrt{p_1 + p_2}/n}{\lambda_{\min}} \right) \), which can be much larger than the bound in Theorem 1 in the high-dimensional region where \( p_1, p_2 \) is comparable to \( n \). This phenomenon also appears in our simulation studies in Section 5.2. The third term corresponds to the discretization error on estimating the function \( f \) from a discrete grid, which is unavoidable even in the noiseless orthogonal-decomposable case (i.e., \( \mathcal{E} = \mathcal{E}_k = \tau = \mu = 0 \)). The last term \( (r - 1)\mu \) corresponds to the coherence of the signal tensor \( \mathcal{X} \): when different singular components are not exactly orthogonal, each of them is not a stationery point even in the noiseless cases. Nevertheless, this term often has negligible rate compared to the other three terms in many scenarios (for example, in the scenario of Proposition 2, \( \mu \lesssim 1/\sqrt{\min(p_1, p_2)} \); see the discussions therein).

It is also noteworthy that when the tensor \( \mathcal{X} \) is rank-1, the estimation error for singular functions does not contain the discretization term \( \eta_n \) while this is not true for \( r > 1 \). The reason is that when there is only one singular component, the “inaccurate” estimation \( \hat{\xi} \) will not introduce additional bias from other singular components when projection is performed in power iteration. On the other hand, the estimation error of \( \hat{\xi} \) will be accumulated to the tabular modes when there are more than one singular components in the signal tensor \( \mathcal{X} \).

Note that the result of Theorem 1 is conditioning on fixed values of the functional remainder term \( \mathcal{E} \). When \( \mathcal{Z} \) is random, we can develop finer upper bounds on \( \mathcal{E} \) as well as the estimation error. For example, if \( \mathcal{Z} \) is a collection of iid Gaussian processes, we have the following result.

**Assumption 5.** Suppose \( Z_{ij} \) are iid mean-zero Gaussian processes with almost sure continuous path and
\[
\mathcal{E} \sup \sup_{t} |Z_{ij}| \leq m, \quad \sup \sup_{t} \text{var}(Z_{ij}) \leq \sigma^2. \tag{17}
\]

Assumption 5 essentially means each functional \( Z_{ij} \) is perturbed within a finite variation/scale that does not grow with tensor dimension \( p_1, p_2 \). Assumption 5 also implies the following upper bound on \( \mathcal{E} \).

**Proposition 3.** Suppose Assumption 5 holds. Then with probability at least \( 1 - e^{-c(p_1+p_2)} \),
\[
\mathcal{E} \leq m + \sigma \sqrt{p_1 + p_2}.
\]

To illustrate our theoretical results, we specifically consider two RKHSs of major interests next. First, if the RKHS kernel \( \mathcal{K} \) has \( d \) nonzero eigenvalues, it follows that \( \epsilon_{\text{fin}} \lesssim \sqrt{d/n} \) (Proposition 6, supplementary materials) and we can obtain the following estimation error upper bound.

**Corollary 1 (Local Convergence: Finite-dimensional RKHS).** Suppose the reproducing kernel \( \mathcal{H} \) of \( \mathcal{H} \) has finitely many, say \( d \), nonzero eigenvalues. Suppose \( \log n \leq d \leq n \), Assumptions 1–5 hold. Then, with probability at least \( 1 - Cn^{-9} - C \log n \cdot e^{-c(p_1+p_2)} \),
\[
\max \{ \text{dist}(\hat{a}, a), \text{dist}(\hat{b}, b), \text{dist}(\hat{\xi}, \xi) \} \\
\lesssim \sqrt{\frac{d}{n} + \frac{m + \sigma \sqrt{p_1 + p_2}}{\lambda_{\min}} + \sqrt{\frac{p_1 + p_2 + d}{n}} \cdot \frac{\tau}{\lambda_{\min}} + (r - 1)\mu},
\]

Note that a different form of the estimation error bound (Equation (10))satisfies this assumption.
Another interesting example is the Sobolev-Hilbert space $W^{α,2}$ defined in (4). If $\mathcal{H} = W^{α,2}$ with constant $α > 1/2$, then $\mathcal{H}$ is an RKHS with eigenvalues $μ_k \propto k^{-2α}$ (see, e.g., Micchelli and Wahba (1981)) and $c_α ≤ n^{-α/(2α + 1)}$ (Proposition 6, supplementary materials). The following estimation error bound holds accordingly:

**Corollary 2 (Local Convergence: Sobolev Space).** Suppose $\mathcal{H} = W^{α,2}$ with $α > 1/2$ and Assumptions 1–5 hold. Then, with probability at least $1 − Cn^{-9} − C \log n \cdot e^{-c/(p_1+p_2)}$,

$$\max\{\text{dist}(\hat{a}, a), \text{dist}(\hat{b}, b), \text{dist}(\hat{ξ}, ξ)\} ≤ n^{-α/(2α + 1)} + \frac{m}{\lambda_{\min}} + \frac{p_1 + p_2}{\lambda_{\min}} + \frac{1}{(r-1)μ} \cdot \frac{\text{dist}(a(0), a), \text{dist}(b(0), b)}{n}. \tag{18}$$

Moreover, if Assumptions 2, 3, and (18) hold, we have the same conclusion as Theorem 1.

**Theorem 2** implies that the initialization condition (Assumption 4) for local convergence (Theorem 1) holds as long as the signal-to-noise ratio (SNR) condition (18) holds.

**5. Numeric Experiments**

In this section, we investigate the performance of the proposed Functional Tensor SVD estimators, which we refer as FTSVD for short, on both synthetic and real datasets. We use the rescaled Bernoulli polynomial as the reproducing kernel throughout the section:

$$\mathbb{K}(x, y) = 1 + k_1(x)k_1(y) + k_2(x)k_2(y) - k_3(|x - y|),$$

where $k_1(x) = x - .5, k_2(x) = (k_1^2(x) - 1/12)/2$, and $k_3(x) = (k_1^3(x) - k_2^2(x)/2 + 7/240)/24$ for any $x \in [0, 1]$. Note that $\mathbb{K}$ is the reproducing kernel for the Hilbert space $W^{2,2}$ (Gu 2013, chap. 2.3.3).

### 5.1. Computation of (9)

We first provide the implementation details for the constraint optimization (9)—a crucial step in updating the singular functions in the proposed algorithm. Denote $k(s) = (\mathbb{K}(s, s_1), \ldots, \mathbb{K}(s, s_n))^T \in \mathbb{R}^n$ for any $s \in [0, 1]$ and $K = [k(s_1), \ldots, k(s_n)] \in \mathbb{R}^{n \times n}$ as the discrete kernel matrix. Then, we can obtain a solution to (9) based on the convex duality and Representer Theorem (Kimeldorf and Wahba 1971).

**Proposition 4.** Let $\tilde{y}^{(l)} := M_3(\tilde{ỹ}) (a^{(l)} \otimes b^{(l)}) \in \mathbb{R}^n$ and let

$$C_{k} = \inf \left\{ μ : μ ≥ 0, \ y^{(l)\top}\left(K + μI\right)^{-1}K(K + μI)^{-1}y^{(l)} \leq (C_{k}\lambda_{\max})^2 \right\}.$$

Then, $\tilde{ỹ}^{(l+1)}(s) = k(s)\tilde{ỹ}$ for all $s \in [0, 1]$ with $β := (K + C_{k}I)^{1/2} \tilde{ỹ}^{(l)} \in \mathbb{R}^{n}$.

Although $C_{k}$ does not have a closed form solution, it can be efficiently computed via binary search since the function $h(μ) := y^{(l)\top}(K + μI)^{-1}K(K + μI)^{-1}y^{(l)}$ is monotonically decreasing. The threshold $(C_{k}\lambda_{\max})^2$ corresponds to the constraint value in (9). Recall $C_{k}$ is the upper bound of the RKHS-norm of singular functions and we use a fixed value $C_{k} = 200$ under the Bernoulli polynomial kernel for all the upcoming experiments. $\lambda_{\max}$ is the largest singular value in the proposed model that is also unknown. Therefore, we apply High-order Orthogonal Iteration (HOI) method (De Lathauwer, De Moor, and Vandewalle 2000) to estimate $\lambda_{\max}$:

$$(\hat{a}, \hat{b}, \hat{v}) = \text{HOOI}(\tilde{ỹ}), \quad \hat{λ}_{\max} = n^{-1/2} \|\tilde{ỹ} \times_1 \hat{a}^T \times_2 \hat{b}^T \times_3 \hat{v}^T\|.$$

Here $(\hat{a}, \hat{b}, \hat{v})$ are the estimates of leading singular vectors of the tabular tensor $\tilde{ỹ}$. Note that in addition to projecting $\tilde{ỹ}$ onto these three directions, we need to divide it by $n^{1/2}$ as $\lambda_{\max}$ is singular value for the functional tensor, not the discretized tabular tensor.
5.2. Simulation Studies

We simulate the $p_1 \times p_2 \times [0, 1]$-dimensional rank-$r$ functional tensor datasets as follows. For each $l \in [r]$, we sample $a_l, b_l$ uniformly from the unit spheres $S^{p_1-1}, S^{p_2-1}$ and generate $\xi_l$ from orthonormal basis functions $\{u_i(s)\}_{i=1}^{10} \subset \mathcal{L}^2([0, 1])$. Following Yuan and Cai (2010) and Wang et al. (2020b), we set $u_1(s) = 1$ and $u_i(s) = \sqrt{2} \cos((i-1)\pi s)$ for $i = 2, \ldots, 10$. We generate $x_{li} \sim \text{Unif}[-1/i, 1/i]$ independently, $\xi_l(\cdot) = \sum_{i=1}^{10} x_{li} u_i(\cdot)$, singular functions $\xi_l = \xi_l^1 / \|\xi_l^1\|_{L^2}$, and the signal tensor $X = \sum_{l=1}^{r} \lambda_l a_l \otimes b_l \otimes \xi_l \in \mathbb{R}^{p_1 \times p_2 \times [0, 1]}$, where $\lambda_l := \lambda_{\text{min}} \cdot (r-l+1)$ and $\lambda_{\text{min}} > 0$ is some pre-specified least singular value. We set the remainder functions $Z_{ijk} = 0$ for the upcoming two experiments and their effects are studied in Appendix A.1, supplementary materials. Finally, we generate a discrete grid $\{s_k\}_{k=1}^{n} = \text{Unif}(0, 1)$ and obtain the observation $\hat{Y} \in \mathbb{R}^{p_1 \times p_2 \times n}$ such that $\hat{Y}_{ijk} = Y_{ijk} + \epsilon_{ijk}$, where $\epsilon_{ijk} \sim \text{Unif}(0, \tau^2)$. For each simulation setting, the estimation errors are calculated by (11) and reported over 100-times repeated experiments.

We start by studying the performance of FTSVD on rank-1 models and set $\lambda_{\text{min}} = 2$, $\tau = 1$, and $\sigma = 0$. We compare our method with the classic method for rank-1 tabular tensor decomposition (which we refer to as CP later) and the functional principle component analysis (FPCA) proposed by Yao, Müller, and Wang (2005). Since CP only yields the estimation on the discrete grid for functional mode, we obtain the whole function estimator by interpolation with minimal $H$-norm. To apply FPCA, we first unfold the original observations tensor $\hat{Y}$ to a matrix $\hat{Y} = M_3(\hat{Y}) \in \mathbb{R}^{n \times (p_1 p_2)}$ and treat each columns of $\hat{Y}$ as a functional sample evaluated at an $n$-grid. The comparison between CP and the proposed FTSVD for dimension $p = p_1 = p_2 \in \{20, 30, 50\}$ and various grid density $n$ is presented in Figure 1. We can see the proposed algorithm has more accurate estimations than the classic CP-decomposition and FPCA in all scenarios on both functional and tabular modes. In particular, on the functional mode, our method is significantly better as we fully use both the functional smoothness and low-dimensional structures for estimation at each step of power iteration.

We then explore the high-dimensional settings with imbalanced tabular dimensions ($p_1 \neq p_2$) and/or multiple components ($r > 1$). We fix $\lambda_{\text{min}} = 8$, $\sigma = \tau = 1$, $n = 30$, and consider the following four specific scenarios: I: $p_1 = 20, p_2 = 500, r = 1$; II: $p_1 = 20, p_2 = 500, r = 2$; III: $p_1 = 100, p_2 = 500, r = 1$; IV: $p_1 = 100, p_2 = 100, r = 3$. FPCA and CP, the two baseline methods, are implemented as follows in the cases of $r > 1$. For FPCA, we unfold the tensor data to matrix similarly as the previous experiment and apply FPCA to obtain the first $r$ eigenfunctions. For CP, we apply the classic power iteration with random initialization proposed by Anandkumar, Ge, and Janzamin (2014) (the number of initializers is set to 20), and the same interpolation procedure is applied to obtain a functional observation as we did in rank-1 situations. The estimation errors of all singular vectors/functions and the corresponding execution time are reported in Table 1. For multiple PCs scenarios, the average estimation errors of all singular vectors/functions (after optimal perturbation) are presented. NAs appear since FPCA does not yield estimates of $a$ or $b$. As we can see, FTSVD achieves better performances on all the four scenarios than CP and FPCA, particularly in singular function estimation. We also observe that our method has significantly better estimation performance for tabular loadings when rank is greater than one. In addition, our method has nearly the same execution time as the classic CP.

![Figure 1. Comparisons among FTSVD, CP, and FPCA under different dimension $p = p_1 = p_2$ and grid density $n$ for rank-1 models. Upper and lower panels plot the estimation errors for singular functions and vectors, respectively. Since FPCA does not yield a tabular loading estimation, we only report its estimation performance on functional mode.](image-url)
We also include additional simulation studies to evaluate the effect of functional perturbation \( Z \) (i.e., \( \sigma > 0 \)) and to compare with state-of-the-art methods in multivariate functional data analysis. See Appendix A.1, supplementary materials for details.

### 5.3. Real Data Analysis

In this section, we apply the proposed method to the longitudinal microbiome study. An example of applying this method to world-wide crop production analysis is postponed to Appendix A.2 in the supplementary materials.

To investigate the change of fecal microbial composition for new-born infants we apply our methods to the Early Childhood Antibiotics and the Microbiome (ECAM) dataset published by Bokulich et al. (2016) (Qiita ID 10249). We consider the 42 infants with multiple fecal microbiome measurements from birth over the first 2 years of life. The infants have fecal microbiome sampled monthly in the first year and bi-monthly in the second year. Among the 42 infants, 24 are vaginally delivered and 18 are Cesarean delivered. A natural question is whether the delivery method affects the composition and development of microbiome in the infants’ gut environment.

We focus on the 50 bacteria genera with nonzero read counts for more than 10% of all the samples. The data can be organized as an order-3 count tensor \( \bar{Y} \in \mathbb{N}^{42 \times 50 \times 19} \), where the three modes represent different subjects (i.e., infants), bacterial genus and sampling time, respectively. To account for the variation in sequencing depth, we transform the count data to the log-normalized data. The singular vectors/functions are learned without the pre-fitted model for prediction. When the label/response is available, one can project the raw high-dimensional functional data on the estimated feature and functional loadings (i.e., \( \hat{\xi}, \hat{\xi}_c \)) to obtain the corresponding subject scores, then apply the pre-fitted model for prediction. When the label/response is available for each subject, a data-driven cross-validation scheme can be applied to select the hyperparameters \( r \) and \( C_b \).

This work focuses on unsupervised dimension reduction for high-order functional data. The methods can be extended to the supervised methods that allows for classification or predictions. Specifically, one can first apply tensor SVD on the training high-order functional data, then train a supervised model with the estimated low-dimensional subjects loadings \( \hat{\alpha} \) as the predictors; when the observations/measurements from out-sample subjects are available, one can project the raw high-dimensional functional data on the estimated feature and functional loadings (i.e., \( \hat{\xi}, \hat{\xi}_c \)) to obtain the corresponding subject scores, then apply the pre-fitted model for prediction. When the label/response is available for each subject, a data-driven cross-validation scheme can be applied to select the hyperparameters \( r \) and \( C_b \).

It is also worth mentioning that there are several recent results that directly solve the tabular tensor decomposition by (accelerated) gradient method with convergence guarantee (Cai et al. 2019; Han, Willett, and Zhang 2021; Tong et al. 2021).

### Table 1. Estimation errors and execution time in high-dimensional settings with imbalanced tabular dimension and multiple ranks.

| Scenarios | Method | \( \text{dist}(\alpha, \hat{\alpha}) \) | \( \text{dist}(b, \hat{b}) \) | \( \text{dist}(\xi, \hat{\xi}) \) | Execution time (s) |
|-----------|--------|-----------------|-----------------|-----------------|------------------|
| I         | FTSVD  | 0.113 (0.023)   | 0.463 (0.040)   | 0.150 (0.087)   | 1.12             |
|           | CP     | 0.114 (0.024)   | 0.465 (0.040)   | 0.516 (0.220)   | 1.08             |
|           | FPCA   | NA              | NA              | 0.370 (0.094)   | 13.84            |
| II        | FTSVD  | 0.112 (0.051)   | 0.376 (0.032)   | 0.308 (0.110)   | 2.28             |
|           | CP     | 0.368 (0.197)   | 0.538 (0.125)   | 0.694 (0.179)   | 37.98            |
|           | FPCA   | NA              | NA              | 0.578 (0.118)   | 17.53            |
| III       | FTSVD  | 0.263 (0.066)   | 0.475 (0.062)   | 0.164 (0.091)   | 2.61             |
|           | CP     | 0.285 (0.150)   | 0.491 (0.113)   | 0.522 (0.238)   | 2.35             |
|           | FPCA   | NA              | NA              | 0.631 (0.152)   | 1.35             |
| IV        | FTSVD  | 0.172 (0.059)   | 0.173 (0.059)   | 0.379 (0.114)   | 2.72             |
|           | CP     | 0.511 (0.143)   | 0.513 (0.140)   | 0.735 (0.137)   | 13.19            |
|           | FPCA   | NA              | NA              | 0.692 (0.076)   | 19.77            |

Note: The standard errors are shown in parentheses. Boldface numbers mean the corresponding method achieves significant smaller estimation errors than the peer methods.
Figure 2. Biplot of three singular vectors on subject mode from ECAM data. Note that each point represents an infant with color indicating the delivery method.

Figure 3. Three singular functions from ECAM data and the aggregated observed trajectory with respect to the second and third singular vector on bacteria mode. Error bands of trajectories are obtained using mean $\pm 1.64 \times \text{(standard error of the mean)}$.

Such methods usually have an advantage over power iteration as they directly aim at the optimization of likelihood, and can successfully remove the effect of incoherence from the final estimation error bound. However, it is unclear that whether and how these gradient-based algorithms and analysis techniques can be applied to the functional setting, which might be another interesting research direction in the future.

Supplementary Materials

The supplementary materials include additional simulation results, real data analysis on worldwide crop production, and proofs of technical results of this paper. The following references are used in the supplementary materials: Adamczak (2008); Adler and Taylor (2009); Bartlett, Bousquet, and Mendelson (2005); Bartlett and Mendelson (2002); Cai, Han, and Zhang (2020); Happ-Kurz (2020); Mendelson (2002); Pisier (1983); Rudelson and Vershynin (2013); Vershynin (2018); Wedin (1972); and Zhang and Zhou (2020).

Funding

The research of R. Han and A. R. Zhang was supported in part by NSF grant CAREER-1944904 and NIH grant R01 GM131399.

References

Adamczak, R. (2008), “A Tail Inequality for Suprema of Unbounded Empirical Processes with Applications to Markov Chains,” Electronic Journal of Probability, 13, 1000–1034. [1005]
Adler, R. J., and Taylor, J. E. (2009), Random Fields and Geometry, New York: Springer. [1005]
Allen, G. I. (2013), “Multi-Way Functional Principal Components Analysis,” in 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 220–223. IEEE. [997]
Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014), “Tensor Decompositions for Learning Latent Variable Models,” The Journal of Machine Learning Research, 15, 2773–2832. [996,999]
Zhang, A., and Han, R. (2019), "Optimal Sparse Singular Value Decomposition for High-Dimensional High-Order Data," *Journal of the American Statistical Association*, 114, 1708–1725. [996,997]

Zhang, A., and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [996,1001]

Zhang, A. R., and Zhou, Y. (2020), "On the Non-asymptotic and Sharp Lower Tail Bounds of Random Variables," *Stat*, 9, e314. [1005]

Zhang, C., Han, R., Zhang, A. R., and Voyles, P. M. (2020), "Denoising Atomic Resolution 4d Scanning Transmission Electron Microscopy Data with Tensor Singular Value Decomposition," *Ultramicroscopy*, 219, 113123. [995]

Zhang, Z., Allen, G. I., Zhu, H., and Dunson, D. (2019), "Tensor Network Factorizations: Relationships between Brain Structural Connectomes and Traits," *Neuroimage*, 197, 330–343. [995]