Statistica Sinica Preprint No: SS-2024-0342	
Title	Functional Tensor Regression
Manuscript ID	SS-2024-0342
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0342
Complete List of Authors	Tongyu Li, Fang Yao and Anru R. Zhang
Corresponding Authors	Fang Yao
E-mails	fyao@math.pku.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Functional Tensor Regression

Tongyu Li, Fang Yao and Anru R. Zhang

Peking University and Duke University

Abstract: Tensor regression has attracted significant attention in statistical research. This study tackles the challenge of handling covariates with smooth varying structures. We introduce a novel framework, termed functional tensor regression, which incorporates both the tensor and functional aspects of the covariate. To address the high dimensionality and functional continuity of the regression coefficient, we employ a low Tucker rank decomposition along with smooth regularization for the functional mode. We develop a functional Riemannian Gauss–Newton algorithm that demonstrates a provable quadratic convergence rate, while the estimation error bound is based on the tensor covariate dimension. Simulations and a neuroimaging analysis illustrate the finite sample performance of the proposed method.

Key words and phrases: Functional tensor, neuroimaging, quadratic convergence, tensor regression.

1. Introduction

Over the past few decades, tensors or multidimensional arrays have increasingly emerged in various scientific applications ranging from genomics (Durham et al., 2018) to recommender systems (Bi et al., 2018) to international relations (Hoff, 2015) to computational imaging (Zhang et al., 2020). Such a structure is highly versatile and allows for the representation of complex data in an organized way. Moving from classic matrix-based methods to tensor-based methods, multi-linear data can be better exploited. Tensor data analysis, built on the cornerstone of tensor representation and approximation (Hackbusch, 2019; Kolda and Bader, 2009), includes low-rank tensor recovery (Gandy et al., 2011; Goldfarb and Qin, 2014), tensor principal component analysis (Lu et al., 2008, 2019), tensor canonical correlation analysis (Kim et al., 2007; Luo et al., 2015), tensor classification (Phan and Cichocki, 2010; Makantasis et al., 2018) and tensor regression (Liu et al., 2021; Guhaniyogi, 2014; Guhaniyogi et al., 2017; Zhou et al., 2021, 2024). As far as tensor regression is concerned, a primary focus is to reveal the dependence of a scalar response on a tensor covariate, while the case of tensor response was also widely studied (Li and Zhang, 2017; Sun and Li, 2017, 2019; Lock, 2018).

In modern applications, tensors with a smoothly varying mode often appear within complex-structured objects. Such a mode represents features like time and spectrum and exhibits a certain level of regularity, whereas the other modes are in a tabular format. This class of tensors is called functional or dynamic (Bi et al., 2021) and has been studied mainly from the perspective of tensor decomposition in recent years. For example, in unsupervised learning for time-varying tensors, Han et al. (2024) addressed singular value decomposition based on the theory of reproducing kernel Hilbert space, Zhang et al. (2021) dealt with CP decomposition in the context of recommender systems, Chen et al. (2022); Han et al. (2024) investigated factor models, and Sun and Li (2019) proposed a clustering method with computational efficiency. Under the supervised setup, Zhou et al. (2023) developed a regression model with a partially observed dynamic tensor as the response, Billio et al. (2024, 2023) tackled Bayesian modeling of time series of multilayer networks via a logistic tensor-on-tensor model, and Chen et al. (2024) studied the recovery of a dynamic tensor that incorporates observations of a matrix evolving smoothly over time.

In this paper, we focus on the regression where the covariate is a functional tensor. On top of tensor decomposition, tensor regression features the difficulty of inverting a covariance tensor. Moreover, in functional tensor regression, we hope to extract the information that accumulates smoothly along the functional mode. Hence our work is closely related to functional data analysis, especially functional linear regression (Ramsay and Silverman, 2005; Ramsay and Dalzell, 1991; Hall and Horowitz, 2007; Crambes et al., 2009). Taking advantage of the functional nature will also facilitate dimension reduction and data representation. Specific examples include:

- *Brain imaging.* Neuroscience is boosted by neuroimaging technologies such as functional magnetic resonance imaging, which often forms functional tensors. It is intriguing to discover the connection between neurodegenerative disorders and brain activity (Zhou et al., 2013).

- *High-dimensional longitudinal data.* In longitudinal microbiome studies, a large number of bacterial taxa are measured from multiple subjects at multiple time points (Han et al., 2024). The influence of micro characteristics on the overall metrics is worth exploring.

- *Multilayer network data.* How a network changes is often of paramount interest

(Zhou and Müller, 2022). The adjacency matrices in multiple snapshots of dynamic networks can be stacked into a tensor (Jing et al., 2021), which provides new insights into network analysis.

To overcome the curse of dimensionality and achieve computational efficiency, most literature on tensor estimation imposes low-rank assumptions on the tensor parameter which often enhance its interpretability at the same time (Han et al., 2022; Kolda and Bader, 2009). However, in several cases like CP decomposition, the tensor parameter may suffer from non-identifiability and even ill-posedness (Zhou et al., 2021). This further complicates the functional tensor parameter whose functional mode should be distinguished. Unlike other tensor models where all modes of the coefficient tensor are treated the same way, our functional tensor regression model emphasizes the evolution along the functional mode and demands correct parameterization. Simply discretizing the functional mode may result in the loss of suitable smoothness in data, which has not been considered for tensor regression. As a consequence, new methods are needed to take into account both the tensor and functional aspects.

Besides the formulation of functional tensor regression, it is important to solve the optimization problem corresponding to the estimation for the tensor coefficient, which is generally non-convex or NP-hard (Hillar and Lim, 2013). A feasible approach is to run local algorithms to refine a warm initialization (Chen et al., 2019; Ahmed et al., 2020), so we propose a functional Riemannian Gauss–Newton method modified from Luo and Zhang (2023). Outperforming various methods, such as gradient descent, alternating minimization, and iterative hard thresholding, the Riemannian Gauss–Newton algorithm achieves a quadratic convergence rate while incurring only a moderate computation cost per iteration. We have established the estimation error bound through theoretical analysis based on a functional analog of the tensor restricted isometry property in the covariates, which states that the norms of certain tensors are well preserved.

We summarize the main contributions of this paper as follows. First, to our best knowledge, this is the first attempt to extend regression to the functional tensor setting. Our method utilizes the temporal modes in tensors, in contrast to the more commonly used tabular approaches. We adopt the Tucker decomposition, and circumvent the issue of non-identifiability by focusing on the overall structure. Within the proposed model, we propose a penalized least squares estimator to highlight the smoothness of changes along the functional mode. Second, when addressing the estimation challenges in functional tensor regression, particularly those involving a significant roughness penalty, we derive a new functional version of Riemannian Gauss–Newton algorithm. We then establish a novel quadratic convergence guarantee for the proposed algorithm. Third, our functional tensor regression framework offers a new solution to characterizing time-varying effects in tensor data analysis. We demonstrate its power through simulated and real data examples. The illustration of our method on neuroimaging reflects the effect over time and strengthens the findings in the classic literature Zhou et al. (2013).

The rest of this paper is organized as follows. Section 2 begins with a review of linear/tensor algebra, and then presents the model and estimation method for functional tensor regression. In Section 3, we establish theoretical results for the estimation error and computational complexity associated with the functional Riemannian Gauss–Newton algorithm described in Section 2. The proposed method is examined by numerical performance in Section 4. Technical proofs are deferred to an online Supplementary Material, while the code and data are made available in a GitHub repository (`https://github.com/kellty/FTReg`).

2. Proposed Methodology

In this section, we introduce the functional tensor regression model together with an estimation method. To that end, we first review some notation that will be used throughout this paper.

2.1 Notation and Preliminaries

Denote the Euclidean norm of any vector $\boldsymbol{v}$ by $\|\boldsymbol{v}\|$. For any matrix $\boldsymbol{M}$, let $\sigma_k(\boldsymbol{M})$ be its kth largest singular value, and $\mathrm{SVD}(\boldsymbol{M})$ and $\mathrm{QR}(\boldsymbol{M})$ be the matrices consisting of the left singular vectors of $\boldsymbol{M}$ and the Q part of the QR decomposition of $\boldsymbol{M}$, respectively. If $\boldsymbol{M}$ is a p-by-q matrix of rank r, then $\mathrm{SVD}(\boldsymbol{M})$ and $\mathrm{QR}(\boldsymbol{M})$ belong to

$$\mathbb{O}_{p,r} = \{\boldsymbol{U} \in \mathbb{R}^{p\times r} : \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_r\},$$

the set of p-by-r matrices with orthonormal columns, where $\boldsymbol{U}^\top$ denotes the transpose of $\boldsymbol{U}$, and $\boldsymbol{I}_r$ denotes the r-by-r identity matrix. For any $\boldsymbol{U} \in \mathbb{O}_{p,r}$, let $\boldsymbol{U}_\perp \in \mathbb{O}_{p,p-r}$

be such that $\boldsymbol{U}^\top \boldsymbol{U}_\perp = 0$, i.e., the columns of $\boldsymbol{U}$ and $\boldsymbol{U}_\perp$ form an orthonormal basis of $\mathbb{R}^p$. The Kronecker product of two matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ is denoted by $\boldsymbol{M}_1 \otimes \boldsymbol{M}_2$; if the (i,j)th entry of $\boldsymbol{M}_1$ is a_{ij}, then the (i,j)th block of $\boldsymbol{M}_1 \otimes \boldsymbol{M}_2$ is $a_{ij}\boldsymbol{M}_2$.

For any tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$, let $[\boldsymbol{\mathcal{T}}]_{j_0,j_1,\dots,j_D}$ be its $(j_0, j_1, \dots, j_D)$th entry. Given two tensors $\boldsymbol{\mathcal{T}}_1, \boldsymbol{\mathcal{T}}_2 \in \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$ of the same order, their Frobenius inner product is defined as

$$\langle \boldsymbol{\mathcal{T}}_1, \boldsymbol{\mathcal{T}}_2 \rangle = \sum_{j_0,j_1,\dots,j_D} [\boldsymbol{\mathcal{T}}_1]_{j_0,j_1,\dots,j_D} [\boldsymbol{\mathcal{T}}_2]_{j_0,j_1,\dots,j_D}.$$

The tensor Frobenius norm is correspondingly defined as $\|\boldsymbol{\mathcal{T}}\|_{\mathrm{F}} = \langle \boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{T}} \rangle^{1/2}$. The d-mode product of $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$ with a matrix $\boldsymbol{M} \in \mathbb{R}^{p_d \times q_d}$ is denoted by $\boldsymbol{\mathcal{T}} \times_d \boldsymbol{M} \in \mathbb{R}^{q_0 \times \cdots \times q_{d-1} \times p_d \times q_{d+1} \times \cdots \times q_D}$ whose entries are

$$[\boldsymbol{\mathcal{T}} \times_d \boldsymbol{M}]_{j_0,\dots,j_D} = \sum_{k=1}^{q_d} [\boldsymbol{\mathcal{T}}]_{j_0,\dots,j_{d-1},k,j_{d+1},\dots,j_D} [\boldsymbol{M}]_{j_d,k}.$$

For convenience, the tensor-matrix product along multiple modes is abbreviated as

$$\boldsymbol{\mathcal{T}} \times_{d=0}^{D} \boldsymbol{M}_d = \boldsymbol{\mathcal{T}} \times_0 \boldsymbol{M}_0 \times_1 \boldsymbol{M}_1 \times \cdots \times_D \boldsymbol{M}_D.$$

Let $\mathscr{M}_d$ be the operation that unfolds tensors $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$ along mode d into matrices $\mathscr{M}_d(\boldsymbol{\mathcal{T}}) \in \mathbb{R}^{q_d \times q_{-d}}$, where $q_{-d} = \prod_{e \neq d} q_e$, which is often termed as matricization. Formally, $[\mathscr{M}_d(\boldsymbol{\mathcal{T}})]_{j_d,k} = [\boldsymbol{\mathcal{T}}]_{j_0,j_1,\dots,j_D}$ if $k = 1 + \sum_{e \neq d}(j_e - 1)\prod_{f<e,\, f \neq d} q_f$. The inverse operation of $\mathscr{M}_d$ is denoted by $\mathscr{T}_d : \mathbb{R}^{q_d \times q_{-d}} \to \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$, called the mode-$d$ tensorization. It can be seen that

$$\mathscr{M}_d(\boldsymbol{\mathcal{T}} \times_0 \boldsymbol{M}_0 \times \cdots \times_D \boldsymbol{M}_D) = \boldsymbol{M}_d \mathscr{M}_d(\boldsymbol{\mathcal{T}})(\boldsymbol{M}_D^\top \otimes \cdots \otimes \boldsymbol{M}_{d+1}^\top \otimes \boldsymbol{M}_{d-1}^\top \otimes \cdots \otimes \boldsymbol{M}_0^\top).$$

The Tucker rank of a tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$ is defined by

$$\operatorname{rank}_{\mathrm{Tuc}}(\boldsymbol{\mathcal{T}}) = (\operatorname{rank} \mathscr{M}_0(\boldsymbol{\mathcal{T}}), \operatorname{rank} \mathscr{M}_1(\boldsymbol{\mathcal{T}}), \ldots, \operatorname{rank} \mathscr{M}_D(\boldsymbol{\mathcal{T}})).$$

Tucker (1966) and De Lathauwer et al. (2000) indicated that, if $\operatorname{rank}_{\mathrm{Tuc}}(\boldsymbol{\mathcal{T}}) = (r_0, r_1, \ldots, r_D)$, then $\boldsymbol{\mathcal{T}}$ admits the higher-order singular value decomposition:

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\mathcal{S}} \times_{d=0}^{D} \boldsymbol{U}_d,$$

where $\boldsymbol{U}_d = \mathrm{SVD}\{\mathscr{M}_d(\boldsymbol{\mathcal{T}})\} \in \mathbb{O}_{q_d, r_d}$ consists of the mode-d singular vectors, and $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_0 \times r_1 \times \cdots \times r_D}$ is the core tensor. Given $\boldsymbol{r} = (r_0, r_1, \ldots, r_D)$, the tensor $\boldsymbol{\mathcal{T}}_{\max(\boldsymbol{r})}$ denotes the best Tucker approximation of $\boldsymbol{\mathcal{T}}$ in the sense that $\boldsymbol{\mathcal{T}}_{\max(\boldsymbol{r})} = \boldsymbol{\mathcal{T}} \times_{d=0}^{D} (\hat{\boldsymbol{U}}_d \hat{\boldsymbol{U}}_d^\top)$ with

$$\|\boldsymbol{\mathcal{T}} \times_{d=0}^{D} (\hat{\boldsymbol{U}}_d \hat{\boldsymbol{U}}_d^\top)\|_{\mathrm{F}} = \max_{\boldsymbol{U}_d \in \mathbb{O}_{q_d, r_d},\, d=0,1,\ldots,D} \left\|\boldsymbol{\mathcal{T}} \times_{d=0}^{D} (\boldsymbol{U}_d \boldsymbol{U}_d^\top)\right\|_{\mathrm{F}}.$$

2.2 Functional Tensor Regression

We consider the regression framework where a scalar response is influenced by a functional tensor covariate, which extends tensor regression to the functional setting. To model the functional mode, we adopt the concepts of the functional linear regression model introduced by Ramsay and Dalzell (1991). Its classical version relates a functional covariate x to the response y via $E(y \mid x) = \int_{\mathbb{T}} x(t)\beta(t)\,\mathrm{d}t$, where β is the unknown coefficient function and $\mathbb{T}$ is the continuous domain on which x and β are defined. This motivates our functional tensor model:

$$y = \int_{\mathbb{T}} \langle \boldsymbol{\mathcal{X}}(t), \boldsymbol{\mathcal{B}}(t) \rangle \,\mathrm{d}t + \varepsilon,$$

where y is a continuous response, $\boldsymbol{\mathcal{X}}(\cdot)$ is a functional tensor covariate, $\boldsymbol{\mathcal{B}}(\cdot)$ is an unknown functional tensor coefficient, and ε is a zero-mean error. The parameter

$\boldsymbol{\mathcal{B}}(\cdot)$ is functional and thus can be evaluated at any time point, which exhibits a smooth time-varying effect. The functions $\boldsymbol{\mathcal{X}}(\cdot)$ and $\boldsymbol{\mathcal{B}}(\cdot)$ are defined on the interval $\mathbb{T}$, take values in $\mathbb{R}^{p_1\times\cdots\times p_D}$, and are required to satisfy some regularity condition. The functional tensor $\boldsymbol{\mathcal{X}}(\cdot)$ admits the Karhunen–Loève decomposition

$$\boldsymbol{\mathcal{X}}(\cdot) = \sum_{k=1}^{\infty} \boldsymbol{\mathit{\Xi}}_k \varphi_k(\cdot), \tag{2.1}$$

where φ_k's form an orthonormal basis of the space $L^2(\mathbb{T})$ of square-integrable functions, and $\boldsymbol{\mathit{\Xi}}_k = \int_{\mathbb{T}} \boldsymbol{\mathcal{X}}(t)\varphi_k(t)\,\mathrm{d}t$ are uncorrelated random tensors. The concrete conditions on $\boldsymbol{\mathcal{X}}(\cdot)$ are listed in Section 3. In addition, the entries of $\boldsymbol{\mathcal{B}}(\cdot)$ lie in some Sobolev space $W^{m,2}(\mathbb{T})$, that is, $\boldsymbol{\mathcal{B}}(\cdot)$ is m-times differentiable and the entries of its mth derivative $\boldsymbol{\mathcal{B}}^{(m)}(\cdot)$ are square-integrable. We assume that the training data $(y_i, \boldsymbol{\mathcal{X}}_i)$, $i = 1, \ldots, n$, consists of n independent copies of $(y, \boldsymbol{\mathcal{X}})$, and that $\boldsymbol{\mathcal{X}}_i$'s are measured only over a discrete grid $t_1 < \cdots < t_{p_0}$ with observational noise, i.e.,

$$\boldsymbol{\mathcal{X}}_{ij} = \boldsymbol{\mathcal{X}}_i(t_j) + \boldsymbol{\mathcal{E}}_{ij}, \tag{2.2}$$

where $\boldsymbol{\mathcal{E}}_{ij}$ is a $p_1 \times \cdots \times p_D$ tensor with entries of zero mean and finite variance. To facilitate global retrieval, we assume further that there exists a constant $C_0 > 0$ for which $C_0^{-1} \le p_0(t_{j+1} - t_j) \le C_0$, $j = 0, 1, \ldots, p_0$, where t_0 and t_{p_0+1} are the endpoints of $\mathbb{T}$. The aligned measurement points t_j's allow us to focus on a tensor structure of the unknown coefficient $\boldsymbol{\mathcal{B}}(\cdot)$, which further leads to an affordable computational burden. It is nontrivial to extend this framework to irregularly spaced observations, and we leave it as future work. In what follows, we write $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ for the vectors of y_i's and ε_i's, respectively, where $\varepsilon_i = y_i - \int_{\mathbb{T}} \langle \boldsymbol{\mathcal{X}}_i(t), \boldsymbol{\mathcal{B}}(t)\rangle\,\mathrm{d}t$.

Based on the above model, our goal is to estimate $\boldsymbol{\mathcal{B}}(\cdot)$. As is shown in Crambes et al. (2009), the influence of $\boldsymbol{\mathcal{X}}(\cdot)$ at each measurement point t_j could be quantified by $\boldsymbol{\mathcal{B}}_j = \boldsymbol{\mathcal{B}}(t_j)$, which leads to the idea of interpolating $\boldsymbol{\mathcal{B}}_j$'s to recover $\boldsymbol{\mathcal{B}}(\cdot)$. Hence, for simplicity, $\boldsymbol{\mathcal{B}}(\cdot)$ is supposed to be an interpolant through $(t_j, \boldsymbol{\mathcal{B}}_j)$, $j = 1, \ldots, p_0$, which turns out to be unique if natural splines are used. See Eubank (1999, Chapter 5) for smoothing splines. Let $\boldsymbol{\psi}(\cdot) = (\psi_1(\cdot), \ldots, \psi_{p_0}(\cdot))^\top$ be a basis of the space of natural splines with order $2m$ and knots $t_1, \ldots, t_{p_0}$. With $\boldsymbol{\Theta} \in \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{p_0 \times p_0}$ being such that $[\boldsymbol{\Theta}]_{j_0, j_1, \ldots, j_D} = [\boldsymbol{\mathcal{B}}_{j_0}]_{j_1, \ldots, j_D}$ and $[\boldsymbol{\Psi}]_{j,k} = \psi_k(t_j)$, the coefficient of interest is

$$\boldsymbol{\mathcal{B}}(\cdot) = \boldsymbol{\Theta} \times_0 \{\boldsymbol{\psi}(\cdot)^\top (\boldsymbol{\Psi}^\top \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^\top\}, \tag{2.3}$$

where the 0-mode product corresponds to spline interpolation in the functional mode. This formulation allows for easy manipulation of unknown parameters, and we can mitigate the curse of dimensionality by imposing additional structure on the tensor $\boldsymbol{\Theta}$; otherwise, the unprocessed vectorization of $\boldsymbol{\Theta}$ would lead to prohibitive computation in an ultra-high dimensional space.

Then, it suffices to solve the estimation problem for $\boldsymbol{\Theta}$ whose zeroth mode is essentially functional. There is a significant difference in functional tensor regression compared with the tabular tensor regression: one must consider time intervals and smoothness to achieve estimation efficiency. For a given value of tuning parameter $\rho > 0$, we obtain an estimator $\widehat{\boldsymbol{\Theta}}$ by minimizing a penalized squared loss $L(\boldsymbol{\Theta}) + \rho J(\boldsymbol{\Theta})$

in some suitable space of $\boldsymbol{\Theta}$. Here

$$\begin{aligned} L(\boldsymbol{\Theta}) &= (2n)^{-1}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p_0}\langle \boldsymbol{\mathcal{X}}_{ij}, \boldsymbol{\mathcal{B}}_j\rangle \Delta t_j\Big)^2 \\ &= (2n)^{-1}\sum_{i=1}^{n}\big(y_i - \langle \boldsymbol{\mathcal{Z}}_i, \boldsymbol{\Theta}\rangle\big)^2 = (2n)^{-1}\|\boldsymbol{y} - \mathscr{X}\boldsymbol{\Theta}\|^2 \end{aligned} \tag{2.4}$$

for some interval sizes $\Delta t_1, \ldots, \Delta t_{p_0} > 0$, where $\boldsymbol{\mathcal{Z}}_i$ is the $p_0 \times p_1 \times \cdots \times p_D$ tensor such that

$$[\boldsymbol{\mathcal{Z}}_i]_{j_0, j_1, \ldots, j_D} = \Delta t_{j_0}[\boldsymbol{\mathcal{X}}_{ij_0}]_{j_1, \ldots, j_D},$$

and $\mathscr{X}$ is the linear map that sends $\boldsymbol{\Theta}$ to the vector $(\langle \boldsymbol{\mathcal{Z}}_1, \boldsymbol{\Theta}\rangle, \ldots, \langle \boldsymbol{\mathcal{Z}}_n, \boldsymbol{\Theta}\rangle)^\top$. We shall choose $\Delta t_j = (t_{j+1} - t_{j-1})/2$, which is the mean of $t_{j+1} - t_j$ and $t_j - t_{j-1}$, and satisfies the quasi-uniform property that

$$C_0^{-1} \le p_0 \Delta t_j \le C_0. \tag{2.5}$$

The penalty term $J(\boldsymbol{\Theta})$ is designed to control the roughness of $\boldsymbol{\mathcal{B}}(\cdot)$ and ensure the existence of a unique penalized least-squares solution, which requires careful treatment in tensor regression (Zhou and Li, 2014). Motivated by the method developed in Crambes et al. (2009), we take

$$J(\boldsymbol{\Theta}) = \int_{\mathbb{T}} \left\|\boldsymbol{\mathcal{B}}^{(m)}(t)\right\|_{\mathrm{F}}^2 \mathrm{d}t + \sum_{j=1}^{p_0} \Delta t_j \left\|\sum_{k=1}^{m} \tilde{\boldsymbol{\mathcal{A}}}_k t_j^{k-1}\right\|_{\mathrm{F}}^2 \tag{2.6}$$

with $(\tilde{\boldsymbol{\mathcal{A}}}_1, \ldots, \tilde{\boldsymbol{\mathcal{A}}}_m)$ being the minimizer of $\sum_{j=1}^{p_0} \Delta t_j \|\boldsymbol{\mathcal{B}}_j - \sum_{k=1}^{m} \boldsymbol{\mathcal{A}}_k t_j^{k-1}\|_{\mathrm{F}}^2$ over all m-tuples of $p_1 \times \cdots \times p_D$ tensors $(\boldsymbol{\mathcal{A}}_1, \ldots, \boldsymbol{\mathcal{A}}_m)$. The first term penalizing roughness appears quite often in regularization approach and reads $\langle \boldsymbol{\Theta} \times_0 \boldsymbol{\Omega}_m, \boldsymbol{\Theta}\rangle$ by (2.3), where

$$\boldsymbol{\Omega}_m = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\Big\{\int_{\mathbb{T}} \boldsymbol{\psi}^{(m)}(t)\boldsymbol{\psi}^{(m)}(t)^\top \,\mathrm{d}t\Big\}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^\top, \tag{2.7}$$

a positive semi-definite p_0-by-p_0 matrix. The second term complements the first term so that $J(\boldsymbol{\Theta})$ is a positive definite quadratic form of $\boldsymbol{\Theta}$. Let $\boldsymbol{G}$ be the p_0-by-m matrix with $[\boldsymbol{G}]_{j,k} = t_j^{k-1}$ and $\boldsymbol{\Delta}$ be the diagonal matrix constructed from Δt_j's. Denote

$$\boldsymbol{P}_m = \boldsymbol{G}(\boldsymbol{G}^\top \boldsymbol{\Delta} \boldsymbol{G})^{-1} \boldsymbol{G}^\top \boldsymbol{\Delta},$$

a (not necessarily orthogonal) projection matrix. Then the second term on the right-hand side of (2.6) can be written as

$$\langle (\boldsymbol{\Theta} \times_0 \boldsymbol{P}_m) \times_0 \boldsymbol{\Delta}, \boldsymbol{\Theta} \times_0 \boldsymbol{P}_m \rangle = \langle \boldsymbol{\Theta} \times_0 (\boldsymbol{\Delta} \boldsymbol{P}_m), \boldsymbol{\Theta} \rangle.$$

Combining this and (2.7), the penalty (2.6) becomes

$$J(\boldsymbol{\Theta}) = \langle \boldsymbol{\Theta} \times_0 \boldsymbol{A}_m, \boldsymbol{\Theta} \rangle = \langle \mathscr{A} \boldsymbol{\Theta}, \boldsymbol{\Theta} \rangle.$$

Here

$$\boldsymbol{A}_m = \boldsymbol{\Omega}_m + \boldsymbol{P}_m^\top \boldsymbol{\Delta} \boldsymbol{P}_m = \boldsymbol{\Omega}_m + \boldsymbol{\Delta} \boldsymbol{P}_m$$

and $\mathscr{A}$ is the linear map sending $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta} \times_0 \boldsymbol{A}_m$. To see the positive definiteness of $\boldsymbol{A}_m$, notice that the null space of $\boldsymbol{\Omega}_m$ is the column space of $\boldsymbol{G}$, and if $\boldsymbol{u} = \boldsymbol{G}\boldsymbol{v} \in \mathbb{R}^{p_0}$ for some $\boldsymbol{v} \in \mathbb{R}^m$,

$$\boldsymbol{u}^\top \boldsymbol{A}_m \boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{\Delta} \boldsymbol{u} \geq (C_0 p_0)^{-1} \|\boldsymbol{u}\|^2$$

by (2.5). On the other hand, the behavior of the eigenvalues of $\boldsymbol{\Omega}_m$ has been well studied (see, e.g., Utreras, 1983), where the smallest nonzero one $\sigma_{p_0-m}(\boldsymbol{\Omega}_m)$ is of order p_0^{-1}. Consequently, there exists some constant $c_m > 0$ such that the smallest eigenvalue of $\boldsymbol{A}_m$ has the lower bound

$$\sigma_{p_0}(\boldsymbol{A}_m) \geq c_m p_0^{-1}. \tag{2.8}$$

We end this subsection with a note that the parameter is no longer infinite-dimensional, but some characteristics of nonparametric methods will still appear. The spline interpolation relies on the observation points, whose number p_0 is not bounded, so we need to trade off between goodness of fit and model parsimony with respect to the functional mode. Meanwhile, the dimensionality of tabular modes plays a prominent role when modeling the whole tensor. Therefore, we are confronted with an exceptionally flexible model and need careful analysis.

2.3 Functional Riemannian Gauss–Newton Algorithm

To fit the functional tensor model, the fidelity to data and the smoothness of the functional parameter are both addressed. Now the loss function can be written in a ridge form:

$$L(\boldsymbol{\Theta}) + \rho J(\boldsymbol{\Theta}) = (2n)^{-1}\|\boldsymbol{y} - \mathscr{X}\boldsymbol{\Theta}\|^2 + \rho\langle \mathscr{A}\boldsymbol{\Theta}, \boldsymbol{\Theta}\rangle, \tag{2.9}$$

incorporating (2.4) and (2.6). The minimization problem associated with (2.9) is fairly complicated due to its extremely high dimensionality, so it is imperative to reduce the parameter size to a manipulable level. To this end, we assume that the true parameter belongs to

$$\mathbb{M}_{\boldsymbol{r}} = \{\boldsymbol{\Theta} \in \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D} : \operatorname{rank}_{\mathrm{Tuc}}(\boldsymbol{\Theta}) = \boldsymbol{r}\}$$

for some $\boldsymbol{r} = (r_0, r_1, \ldots, r_D)$. Note that Tucker decomposition is flexible for allowing different numbers of factors along each mode. The flexibility in selecting different ranks for various tensor modes is advantageous when the tensor data are dimensionally skewed, a scenario that commonly appears in neuroimaging data (Li et al., 2018).

The set $\mathbb{M}_{\boldsymbol{r}}$ is a smooth manifold of dimension $\prod_{d=0}^{D} r_d + \sum_{d=0}^{D} r_d(p_d - r_d)$ embedded into $\mathbb{R}^{p_0\times p_1\times\cdots\times p_D}$ (Koch and Lubich, 2010; Uschmajew and Vandereycken, 2013). Note that the intrinsic dimension also has important implications for the theoretical analysis, especially with respect to the requirement on the sample size n. Riemannian optimization techniques shed light on the problem of minimizing (2.9) on $\mathbb{M}_{\boldsymbol{r}}$; see Absil et al. (2009); Boumal (2023) for an introduction. In a Riemannian optimization procedure, an iteration step is typically carried out by updating the point on the tangent space and then retracting it to the manifold. Next, we derive a functional Riemannian Gauss–Newton scheme. Compared to Luo and Zhang (2023), the novelty of our work on functional tensor regression lies in the appropriate treatment of the penalty term, which enables fast computation as established in Proposition 1 below.

We start by describing the tangent space of $\mathbb{M}_{\boldsymbol{r}}$. Let the Tucker decomposition of $\boldsymbol{\Theta} \in \mathbb{M}_{\boldsymbol{r}}$ be $\boldsymbol{\Theta} = \boldsymbol{\mathcal{S}} \times_{d=0}^{D} \boldsymbol{U}_d$ where $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{r_0\times r_1\times\cdots\times r_D}$ and $\boldsymbol{U}_d \in \mathbb{O}_{p_d,r_d}$, and let

$$\boldsymbol{V}_d = \mathrm{QR}\{\mathscr{M}_d(\boldsymbol{\mathcal{S}})^\top\} \in \mathbb{O}_{r_{-d},r_d}$$

where $r_{-d} = \prod_{e\neq d} r_e$. By Koch and Lubich (2010, Theorem 2.1), the elements of the tangent space $\mathrm{T}_{\boldsymbol{\Theta}}\mathbb{M}_{\boldsymbol{r}}$ of $\mathbb{M}_{\boldsymbol{r}}$ at $\boldsymbol{\Theta}$ are

$$\boldsymbol{\mathcal{C}} \times_{d=0}^{D} \boldsymbol{U}_d + \sum_{d=0}^{D} \mathscr{T}_d(\boldsymbol{D}_d\boldsymbol{V}_d^\top) \times_d \boldsymbol{U}_{d\perp} \times_{e\neq d} \boldsymbol{U}_e = \boldsymbol{\mathcal{C}} \times_{d=0}^{D} \boldsymbol{U}_d + \sum_{d=0}^{D} \mathscr{T}_d(\boldsymbol{U}_{d\perp}\boldsymbol{D}_d\boldsymbol{W}_d^\top),$$

$\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_0\times r_1\times\cdots\times r_D}$, $\boldsymbol{D}_d \in \mathbb{R}^{(p_d-r_d)\times r_d}$, $d = 0, 1, \ldots, D$. Here $\boldsymbol{W}_d$ is defined to be

$$\boldsymbol{W}_d = (\boldsymbol{U}_D \otimes \cdots \otimes \boldsymbol{U}_{d+1} \otimes \boldsymbol{U}_{d-1} \otimes \cdots \otimes \boldsymbol{U}_0)\boldsymbol{V}_d, \tag{2.10}$$

which corresponds to the row space of $\mathscr{M}_d(\boldsymbol{\Theta})$. Thus, the tangent space $\mathrm{T}_{\boldsymbol{\Theta}}\mathbb{M}_{\boldsymbol{r}}$ can be parameterized by

$$\mathbb{D} = \mathbb{R}^{r_0 \times r_1 \times \cdots \times r_D} \times \mathbb{R}^{(p_0 - r_0) \times r_0} \times \mathbb{R}^{(p_1 - r_1) \times r_1} \times \cdots \times \mathbb{R}^{(p_D - r_D) \times r_D}.$$

In conjunction with Koch and Lubich (2010, Lemma 3.1), Luo and Zhang (2023) showed that the projection operator $\mathscr{P}_{\boldsymbol{\Theta}}$ from $\mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D}$ onto $\mathrm{T}_{\boldsymbol{\Theta}}\mathbb{M}_{\boldsymbol{r}}$ can be decomposed into $\mathscr{P}_{\boldsymbol{\Theta}} = \mathscr{R}_{\boldsymbol{\Theta}}\mathscr{R}^*_{\boldsymbol{\Theta}}$, where

$$\begin{aligned} \mathscr{P}_{\boldsymbol{\Theta}} : & \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D} \to \mathrm{T}_{\boldsymbol{\Theta}}\mathbb{M}_{\boldsymbol{r}} \subset \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D} \\ & \boldsymbol{\Upsilon} \mapsto \boldsymbol{\Upsilon} \times_{d=0}^{D} (\boldsymbol{U}_d \boldsymbol{U}_d^\top) + \sum_{d=0}^{D} \mathscr{T}_d\{\boldsymbol{U}_{d\perp}\boldsymbol{U}_{d\perp}^\top \mathscr{M}_d(\boldsymbol{\Upsilon})\boldsymbol{W}_d\boldsymbol{W}_d^\top\}, \end{aligned} \tag{2.11}$$

$$\mathscr{R}_{\boldsymbol{\Theta}} : \big(\boldsymbol{\mathcal{C}}, (\boldsymbol{D}_d)_{d=0}^{D}\big) \in \mathbb{D} \mapsto \boldsymbol{\mathcal{C}} \times_{d=0}^{D} \boldsymbol{U}_d + \sum_{d=0}^{D} \mathscr{T}_d(\boldsymbol{U}_{d\perp}\boldsymbol{D}_d\boldsymbol{W}_d^\top) \in \mathrm{T}_{\boldsymbol{\Theta}}\mathbb{M}_{\boldsymbol{r}}, \tag{2.12}$$

$$\mathscr{R}^*_{\boldsymbol{\Theta}} : \boldsymbol{\Upsilon} \in \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D} \mapsto \big(\boldsymbol{\Upsilon} \times_{d=0}^{D} \boldsymbol{U}_d^\top, \{\boldsymbol{U}_{d\perp}^\top \mathscr{M}_d(\boldsymbol{\Upsilon})\boldsymbol{W}_d\}_{d=0}^{D}\big) \in \mathbb{D}. \tag{2.13}$$

The linear operators $\mathscr{R}_{\boldsymbol{\Theta}}$ and $\mathscr{R}^*_{\boldsymbol{\Theta}}$ represent extension and contraction, respectively. The image of $\mathscr{R}_{\boldsymbol{\Theta}}$ contains generic elements of $\mathrm{T}_{\boldsymbol{\Theta}}\mathbb{M}_{\boldsymbol{r}}$, and $\mathscr{R}^*_{\boldsymbol{\Theta}}$ is constructed such that $\mathscr{R}_{\boldsymbol{\Theta}}\mathscr{R}^*_{\boldsymbol{\Theta}} = \mathscr{P}_{\boldsymbol{\Theta}}$.

Given the current iterate $\boldsymbol{\Theta}^k$, the objective function (2.9) evaluated at $\mathscr{R}_{\boldsymbol{\Theta}^k}\big(\boldsymbol{\mathcal{C}}, (\boldsymbol{D}_d)_{d=0}^{D}\big)$ in place of $\boldsymbol{\Theta}$ is

$$(2n)^{-1}\big\|\boldsymbol{y} - \mathscr{X}\mathscr{R}_{\boldsymbol{\Theta}^k}\big(\boldsymbol{\mathcal{C}}, (\boldsymbol{D}_d)_{d=0}^{D}\big)\big\|^2 + \rho\langle \mathscr{A}\mathscr{R}_{\boldsymbol{\Theta}^k}\big(\boldsymbol{\mathcal{C}}, (\boldsymbol{D}_d)_{d=0}^{D}\big), \mathscr{R}_{\boldsymbol{\Theta}^k}\big(\boldsymbol{\mathcal{C}}, (\boldsymbol{D}_d)_{d=0}^{D}\big)\rangle,$$

where the number of parameters is equal to $\dim \mathbb{M}_{\boldsymbol{r}} = \prod_{d=0}^{D} r_d + \sum_{d=0}^{D} r_d(p_d - r_d)$ and becomes computationally feasible. The solution minimizing it is explicitly

$$\breve{\boldsymbol{\Theta}}^{k+1} = \mathscr{R}_{\boldsymbol{\Theta}^k}(\mathscr{R}^*_{\boldsymbol{\Theta}^k}\mathscr{X}^*\mathscr{X}\mathscr{R}_{\boldsymbol{\Theta}^k} + n\rho\mathscr{R}^*_{\boldsymbol{\Theta}^k}\mathscr{A}\mathscr{R}_{\boldsymbol{\Theta}^k})^{-1}\mathscr{R}^*_{\boldsymbol{\Theta}^k}\mathscr{X}^*\boldsymbol{y}, \tag{2.14}$$

where $\mathscr{Z}^*$ is the adjoint operator of $\mathscr{Z}$, i.e., $\boldsymbol{y} \mapsto \sum_{i=1}^n y_i \boldsymbol{\mathcal{Z}}_i$. The regularization ensures that there is no problem with matrix invertibility when $\rho > 0$. For expressing $\mathscr{R}^*_{\boldsymbol{\Theta}^k} \mathscr{A} \mathscr{R}_{\boldsymbol{\Theta}^k}$, we have the following proposition.

Proposition 1. *Let* $\mathscr{R}_{\boldsymbol{\Theta}}$ *and* $\mathscr{R}^*_{\boldsymbol{\Theta}}$ *be defined in* (2.12) *and* (2.13)*, and* $\mathscr{A}$ *be the linear map* $\boldsymbol{\Theta} \mapsto \boldsymbol{\Theta} \times_0 \boldsymbol{A}$. *Then for* $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_0 \times r_1 \times \cdots \times r_D}$, $\mathbf{D}_d \in \mathbb{R}^{(p_d - r_d) \times r_d}$, $d = 0, 1, \ldots, D$,

$$\begin{aligned}\mathscr{R}^*_{\boldsymbol{\Theta}} \mathscr{A} \mathscr{R}_{\boldsymbol{\Theta}}\big(\boldsymbol{\mathcal{C}}, (\boldsymbol{D}_d)_{d=0}^D\big) = \big(&\boldsymbol{\mathcal{C}} \times_0 (\boldsymbol{U}_0^\top \boldsymbol{A} \boldsymbol{U}_0) + \mathscr{T}_0(\boldsymbol{D}_0 \boldsymbol{V}_0^\top) \times_0 (\boldsymbol{U}_0^\top \boldsymbol{A} \boldsymbol{U}_{0\perp}),\\ &\boldsymbol{U}_{0\perp}^\top \boldsymbol{A} \boldsymbol{U}_0 \mathscr{M}_0(\boldsymbol{\mathcal{C}}) \boldsymbol{V}_0 + \boldsymbol{U}_{0\perp}^\top \boldsymbol{A} \boldsymbol{U}_{0\perp} \boldsymbol{D}_0,\\ &\big[\mathscr{M}_d\{\mathscr{T}_d(\boldsymbol{D}_d \boldsymbol{V}_d^\top) \times_0 (\boldsymbol{U}_0^\top \boldsymbol{A} \boldsymbol{U}_0)\} \boldsymbol{V}_d\big]_{d=1}^D\big).\end{aligned}$$

Then we map $\check{\boldsymbol{\Theta}}^{k+1} \in \mathrm{T}_{\boldsymbol{\Theta}^k} \mathbb{M}_{\boldsymbol{r}}$ in (2.14) back to the manifold $\mathbb{M}_{\boldsymbol{r}}$. An ideal choice is the truncated higher-order singular value decomposition (T-HOSVD) which constructs lower Tucker rank approximations; see Vannieuwenhoven et al. (2012). The detailed procedure of T-HOSVD is given in Appendix A. Denote the T-HOSVD operation by $\mathcal{H}_{\boldsymbol{r}} : \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D} \to \mathbb{M}_{\boldsymbol{r}}$. It is known that $\mathcal{H}_{\boldsymbol{r}}$ satisfies the quasi-projection property (Hackbusch, 2019, Theorem 10.2)

$$\|\boldsymbol{\varUpsilon} - \mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\varUpsilon})\|_{\mathrm{F}} \leq (D+1)^{1/2} \|\boldsymbol{\varUpsilon} - \boldsymbol{\varUpsilon}'\|_{\mathrm{F}}, \quad \boldsymbol{\varUpsilon} \in \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D},\ \boldsymbol{\varUpsilon}' \in \mathbb{M}_{\boldsymbol{r}}. \tag{2.15}$$

In view of this, we update the solution to (2.9) by

$$\boldsymbol{\Theta}^{k+1} = \mathcal{H}_{\boldsymbol{r}}(\check{\boldsymbol{\Theta}}^{k+1}). \tag{2.16}$$

The above functional Riemannian Gauss–Newton steps are summarized in Algorithm 1 below. Note that the sample size n and parameter dimension $p_0, p_1, \ldots, p_D$ are given by data, while the rank $\boldsymbol{r}$ and the tuning parameter ρ needs selection, for which we suggest a generalized cross-validation detailed in Section 4.

Algorithm 1: The functional Riemannian Gauss–Newton scheme for minimizing (2.9)

```
Input: response vector $\boldsymbol{y} \in \mathbb{R}^n$, weighted covariate tensor
       $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{n \times p_0 \times p_1 \times \cdots \times p_D}$, Tucker rank $\boldsymbol{r} = (r_0, r_1, \ldots, r_D)$, multiple of
       penalty matrix $\rho \boldsymbol{A}_m$, iteration number $K$
begin
    Initialize $\boldsymbol{\Theta}^0 = \boldsymbol{\mathcal{S}}^0 \times_{d=0}^{D} \boldsymbol{U}_d^0$;
    for $k = 0, 1, \ldots, K-1$ do
        for $d = 0, 1, \ldots, D$ do
            Calculate $\boldsymbol{V}_d^k = \mathrm{QR}\{\mathscr{M}_d(\boldsymbol{\mathcal{S}}^k)^\top\}$ and $\boldsymbol{W}_d^k$ as in (2.10);
        end
        Obtain $\check{\boldsymbol{\Theta}}^{k+1}$ using (2.14);
        Update $\boldsymbol{\Theta}^{k+1} = \boldsymbol{\mathcal{S}}^{k+1} \times_{d=0}^{D} \boldsymbol{U}_d^{k+1} = \mathcal{H}_{\boldsymbol{r}}(\check{\boldsymbol{\Theta}}^{k+1})$ where $\mathcal{H}_{\boldsymbol{r}}$ is T-HOSVD;
    end
end
Output: $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^K$
```

3. Theoretical Results

In this section, we analyze the error bound of the functional Riemannian Gauss–Newton algorithm in Subsection 2.3. It will be shown that the convergence rate is quadratic which outperforms the commonly used first-order methods. We consider varying samples sizes n and dimensions p_d, r_d, $d = 0, 1, \ldots, D$. Various constants are introduced, which we list in Appendix B for easy reference.

Based on (2.1) and (2.2), we require the following assumptions to guarantee the efficiency of estimation.

Assumption 1. For each k, the tensor $\boldsymbol{\mathit{\Xi}}_k$ has entries with zero mean. Moreover, there exist some constants $1 < a, A < \infty$ such that

$$A^{-1}k^{-a}\|\boldsymbol{\mathit{\Upsilon}}\|_{\mathrm{F}}^2 \le E(\langle \boldsymbol{\mathit{\Xi}}_k, \boldsymbol{\mathit{\Upsilon}}\rangle^2) \le Ak^{-a}\|\boldsymbol{\mathit{\Upsilon}}\|_{\mathrm{F}}^2, \quad \boldsymbol{\mathit{\Upsilon}} \in \mathbb{R}^{p_1\times\cdots\times p_D}.$$

Assumption 2. Let $\boldsymbol{\Phi}_\ell$ be the p_0-by-p_0 matrix with

$$[\boldsymbol{\Phi}_\ell]_{j,k} = (\Delta t_j)^{1/2}\varphi_{k+(\ell-1)p_0}(t_j), \quad j,k = 1,\ldots,p_0.$$

There exists some constant $0 < C_\varphi < \infty$ such that for $\ell = 1, 2, \ldots$, the spectral norm

$$\sigma_1(\boldsymbol{\Phi}_\ell) \le C_\varphi.$$

Assumption 3. The random tensors $\boldsymbol{\mathcal{E}}_{ij}$ in (2.2) are independent and have uncorrelated entries with mean zero and variance $\sigma_X^2 < \infty$.

Assumption 1 imposes smoothness in terms of the decay of covariance operators of $\boldsymbol{\mathit{\Xi}}_k$, which in dimension one coincides with the smoothness condition stated in the literature on functional data analysis (Cai and Hall, 2006; Hall and Horowitz, 2007). Assumption 2 implies that the eigenfunctions capture signals with bounded transformations and prevent the functional pattern from being too wild, that is, the weighted sampling matrices $\boldsymbol{\Phi}_\ell$ have bounded operator norms. Indeed, $\boldsymbol{\Phi}_\ell$'s turn out to be orthogonal matrices when, for instance, $\varphi_k(t) = \sin(2k\pi t)$ and $t_j = (j-1/2)/p_0$.

Then as illustrated by (3.17) and (3.19) in Lemma 1, we are able to offer a substitute for the technical tensor restricted isometry property in Luo and Zhang (2023); see also Remark 2 for the requirement of the sample size n when the considered tensors admit low Tucker rank.

Lemma 1. *Suppose that Assumptions 1–3 hold. For any $\boldsymbol{\Upsilon} \in \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D}$, the probability that*

$$n^{-1}\|\mathscr{X}\boldsymbol{\Upsilon}\|^2 \leq R_u p_0^{-1}\|\boldsymbol{\Upsilon}\|_{\mathrm{F}}^2 \tag{3.17}$$

tends to one as $n \to \infty$, where $R_u = (a-1)^{-1}(2a-1)AC_0C_\varphi + C_0^2\sigma_X^2$. Furthermore, writing $\boldsymbol{\Phi}_0$ for the p_0-by-r_0 matrix such that $[\boldsymbol{\Phi}_0]_{j,k} = (\Delta t_j)^{1/2}\varphi_k(t_j)$, if for some $c > 0$,

$$\|\boldsymbol{\Upsilon} \times_0 \boldsymbol{\Phi}_0^\top\|_{\mathrm{F}} \geq c\|\boldsymbol{\Upsilon}\|_{\mathrm{F}}, \tag{3.18}$$

then the probability that

$$n^{-1}\|\mathscr{X}\boldsymbol{\Upsilon}\|^2 \geq R_l p_0^{-1}\|\boldsymbol{\Upsilon}\|_{\mathrm{F}}^2 \tag{3.19}$$

tends to one as $n \to \infty$, where $R_l = c^2/(AC_0r_0^a)$.

Compared with the tensor restricted isometry property, in Lemma 1 we relax the requirement that $R_l \in (0,1)$ and $R_u \in (1,2)$. Condition (3.18) means that the functional mode of $\boldsymbol{\Upsilon}$ is well accounted for by the leading r_0 eigenfunctions, which is reasonable in the estimation procedure. We make this more precise in the following lemma.

Lemma 2. *For $\boldsymbol{\Upsilon} \in \mathbb{R}^{p_0 \times p_1 \times \cdots \times p_D}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{p_0 \times r_0}$, it holds that*

$$\|\boldsymbol{\Upsilon} \times_0 \boldsymbol{\Phi}^\top\|_{\mathrm{F}} \geq \left\{\sigma_{r_0}(\boldsymbol{\Phi}^\top\boldsymbol{U}_0)\alpha - \sigma_1(\boldsymbol{\Phi}^\top\boldsymbol{U}_{0\perp})(1-\alpha^2)^{1/2}\right\}\|\boldsymbol{\Upsilon}\|_{\mathrm{F}},$$

where $\boldsymbol{U}_0 \in \mathbb{O}_{p_0,r_0}$ *consists of the leading* r_0 *columns of* $\mathrm{SVD}\{\mathscr{M}_0(\boldsymbol{\Upsilon})\}$, *and* $\alpha \in (0,1]$ *such that* $\alpha\|\boldsymbol{\Upsilon}\|_{\mathrm{F}} = \left\|\boldsymbol{\Upsilon} \times_0 (\boldsymbol{U}_0\boldsymbol{U}_0^\top)\right\|_{\mathrm{F}}$.

With the help of Lemma 1, we can establish the following deterministic convergence theory. Let

$$C_m = \int_{\mathbb{T}} \left(\|\boldsymbol{\mathcal{B}}(t)\|_{\mathrm{F}}^2 + \|\boldsymbol{\mathcal{B}}^{(m)}(t)\|_{\mathrm{F}}^2\right) \mathrm{d}t \tag{3.20}$$

and

$$\zeta_m = \sup\left\{\zeta \geq 0 : \sup_{\Upsilon \in \mathbb{M}_{\boldsymbol{r}} : \|\boldsymbol{\Upsilon}-\boldsymbol{\Theta}\|_{\mathrm{F}} \leq \zeta} \langle \mathscr{A}\mathscr{P}_{\boldsymbol{\Upsilon}}\boldsymbol{\Theta}, \mathscr{P}_{\boldsymbol{\Upsilon}}\boldsymbol{\Theta}\rangle \leq 2C_m\right\}. \tag{3.21}$$

Theorem 1. *Recall the updates* (2.14) *and* (2.16) *in the functional Riemannian Gauss–Newton algorithm for minimizing* (2.9)*. If* (3.17) *holds for* $\boldsymbol{\Upsilon} = \boldsymbol{\Theta} - \mathscr{P}_{\boldsymbol{\Theta}^k}\boldsymbol{\Theta}$, (3.19) *holds for* $\boldsymbol{\Upsilon} = \check{\boldsymbol{\Theta}}^{k+1} - \boldsymbol{\Theta}$, *and* $\left\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\right\|_{\mathrm{F}} \leq \zeta_m$, *then*

$$\begin{aligned}\|\boldsymbol{\Theta}^{k+1} - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq{}& \{(D+1)^{1/2}+1\}(D+1)8^{1/2}R_u^{1/2}R_l^{-1/2}\lambda_{\min}^{-1}\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\|_{\mathrm{F}}^2 \\ &+ \{(D+1)^{1/2}+1\}R_l^{-1/2}\eta,\end{aligned} \tag{3.22}$$

where $\lambda_{\min} = \min_{0\leq d\leq D}\sigma_{r_d}\{\mathscr{M}_d(\boldsymbol{\Theta})\}$, *and* $\eta > 0$ *is defined by*

$$\eta^2 = 12p_0n^{-1}\|\boldsymbol{\delta}\|^2 + 16\rho p_0 C_m + 2c_m^{-1}p_0^2n^{-2}\rho^{-1}\left\|(\mathscr{X}^*\boldsymbol{\varepsilon})_{\max(2\boldsymbol{r})}\right\|_{\mathrm{F}}^2. \tag{3.23}$$

Here $\boldsymbol{\delta}$ *is the vector of approximation errors*

$$\delta_i = \int_{\mathbb{T}} \langle \boldsymbol{\mathcal{X}}_i(t), \boldsymbol{\mathcal{B}}(t)\rangle \,\mathrm{d}t - \sum_{j=1}^{p_0}\langle \boldsymbol{\mathcal{X}}_i(t_j), \boldsymbol{\mathcal{B}}_j\rangle \Delta t_j, \quad i = 1, \ldots, n,$$

and c_m *is given in* (2.8).

The error bound (3.22) ensures the functional Riemannian Gauss–Newton iterate to converge quadratically to the ball centered at $\boldsymbol{\Theta}$ of radius $\mathcal{O}(\eta)$. The additional term η^2 includes three parts that result from approximation (first summand), regularization (second), and observational noise (third), respectively. We defer the investigation of η to Theorem 2. Looking closer at the right-hand side of (3.22), one can partition the process of convergence into two phases.

Corollary 1. *Under the assumptions in Theorem 1,*

if $\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\|_{\mathrm{F}} \geq 8^{-1/4} R_u^{-1/4} (D+1)^{-1/2} \lambda_{\min}^{1/2} \eta^{1/2}$, *then*

$$\|\boldsymbol{\Theta}^{k+1} - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq 2\{(D+1)^{1/2}+1\}(D+1)8^{1/2} R_u^{1/2} R_l^{-1/2} \lambda_{\min}^{-1} \|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\|_{\mathrm{F}}^2,$$

which renders quadratic convergence; and

if $\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq 8^{-1/4} R_u^{-1/4} (D+1)^{-1/2} \lambda_{\min}^{1/2} \eta^{1/2}$, *then*

$$\|\boldsymbol{\Theta}^{k+1} - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq 2\{(D+1)^{1/2}+1\} R_l^{-1/2} \eta,$$

which reflects the eventual estimation error.

Consequently, if $\|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq 4^{-1}\{(D+1)^{1/2}+1\}^{-1}(D+1)^{-1} 8^{-1/2} R_u^{-1/2} R_l^{1/2} \lambda_{\min}$, *then*

$$\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq 2^{-2^k} \|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}\|_{\mathrm{F}}$$

for $k \leq K = \inf\{k : \left\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\right\|_{\mathrm{F}} \leq 8^{-1/4} R_u^{-1/4} (D+1)^{-1/2} \lambda_{\min}^{1/2} \eta^{1/2}\}$; *and*

if $8^{-1/4} R_u^{-1/4} (D+1)^{-1/2} \lambda_{\min}^{1/2} \eta^{1/2} \geq 2\{(D+1)^{1/2}+1\} R_l^{-1/2} \eta$, *then*

$$\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\|_{\mathrm{F}} \leq 2\{(D+1)^{1/2}+1\} R_l^{-1/2} \eta$$

for $k > K$. *The above-defined* K *satisfies that*

$$K \leq \lceil \log \max\{1, 2^{-1} \log 16^{-1}\{(D+1)^{1/2}+1\}^{-2}(D+1)^{-1} 8^{-1/2} R_u^{-1/2} R_l \lambda_{\min} \eta^{-1}\}\rceil.$$

Corollary 1 implies that we only need $\mathcal{O}(\log\log \eta^{-1})$ iterations to achieve the estimation error bound. The magnitude of η plays an essential role in both the estimation error bound and the required number of iterations, and we show it in Theorem 2. To characterize the approximation error, we invoke the concept of Hölder continuity. A function $\boldsymbol{\mathcal{X}} : \mathbb{T} \to \mathbb{R}^{p_1\times\cdots\times p_D}$ is said to be κ-Hölder continuous if $\|\boldsymbol{\mathcal{X}}(t) - \boldsymbol{\mathcal{X}}(s)\|_{\mathrm{F}}/|t-s|^{\kappa}$ is uniformly bounded for $t, s \in \mathbb{T}$ with $t \neq s$.

Theorem 2. *Suppose that Assumptions 1–3 hold and that $\boldsymbol{\mathcal{X}}(\cdot)$ is almost surely κ-Hölder continuous for some $0 < \kappa \le 1$. If ε follows the normal distribution $\mathcal{N}(0, \sigma_y^2)$, then η defined by (3.23) satisfies that*

$$\eta^2 = \mathcal{O}_{\mathrm{pr}}\left\{p_0^{1-2\kappa}C_m + \rho p_0 C_m + (n\rho)^{-1}p_0\Big(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d\Big)\right\}.$$

If $\rho \asymp \big\{n^{-1}\big(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d\big)/C_m\big\}^{1/2}$, then the bound can be reduced to

$$\eta^2 = \mathcal{O}_{\mathrm{pr}}\left\{p_0^{1-2\kappa}C_m + n^{-1/2}p_0\Big(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d\Big)^{1/2} C_m^{1/2}\right\}. \tag{3.24}$$

The normality of ε is required in Theorem 2 so that $\big\|(\mathscr{X}^*\boldsymbol{\varepsilon})_{\max(2\boldsymbol{r})}\big\|_{\mathrm{F}}$ has a light-tailed distribution, given Lemma 1 that controls the operator norm of $\mathscr{X}$.

Due to (2.3), the quantity C_m defined in (3.20) has the same order with $p_0^{-1}\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2$. By Luo and Zhang (2023, Lemma 12),

$$\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2 = \mathcal{O}_{\mathrm{pr}}\Big(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d\Big)$$

if $\boldsymbol{\Theta} = \boldsymbol{\Upsilon}_{\max(\boldsymbol{r})}$ for some $\boldsymbol{\Upsilon} \in \mathbb{R}^{p_0\times p_1\times\cdots\times p_D}$ with i.i.d. $\mathcal{N}(0,1)$ entries. In such a case, combining Corollary 1 and Theorem 2, we conclude that the final estimator $\widehat{\boldsymbol{\Theta}}$ has the following error upper bound

$$\left\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\right\|_{\mathrm{F}}^{2}=\mathcal{O}\big(\eta^{2}\big)=\mathcal{O}_{\mathrm{pr}}\Big\{\big(p_0^{-2\kappa}+n^{-1/2}p_0^{1/2}\big)\Big(\sum_{d=0}^{D}p_d r_d+\prod_{d=0}^{D}r_d\Big)\Big\} \tag{3.25}$$

when the tuning parameter $\rho \asymp (p_0/n)^{1/2}$. Based on (2.3), let $\widehat{\boldsymbol{\mathcal{B}}}(\cdot)$ be the plug-in estimator using $\widehat{\boldsymbol{\Theta}}$. It is clear that

$$\int_{\mathbb{T}}\left\|\widehat{\boldsymbol{\mathcal{B}}}(t)-\boldsymbol{\mathcal{B}}(t)\right\|_{\mathrm{F}}^{2}\mathrm{d}t=\Big\langle\big(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\big)\times_0\boldsymbol{\Omega}_0,\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\Big\rangle,$$

where $\boldsymbol{\Omega}_0$ is defined similarly as (2.7). Since $\boldsymbol{\Omega}_0 \asymp p_0^{-1}\boldsymbol{I}_{p_0}$ as $p_0 \to \infty$, we deduce from (3.25) that

$$\int_{\mathbb{T}}\left\|\widehat{\boldsymbol{\mathcal{B}}}(t)-\boldsymbol{\mathcal{B}}(t)\right\|_{\mathrm{F}}^{2}\mathrm{d}t=\mathcal{O}_{\mathrm{pr}}\Big[\big\{p_0^{-2\kappa-1}+(np_0)^{-1/2}\big\}\Big(\sum_{d=0}^{D}p_d r_d+\prod_{d=0}^{D}r_d\Big)\Big].$$

Furthermore, given $\int_{\mathbb{T}}\|\boldsymbol{\mathcal{B}}(t)\|_{\mathrm{F}}^{2}\,\mathrm{d}t \asymp p_0^{-1}\|\boldsymbol{\Theta}\|_{\mathrm{F}}^{2} \asymp p_0^{-1}\big(\sum_{d=0}^{D}p_d r_d+\prod_{d=0}^{D}r_d\big)$, it is seen that

$$\int_{\mathbb{T}}\left\|\widehat{\boldsymbol{\mathcal{B}}}(t)-\boldsymbol{\mathcal{B}}(t)\right\|_{\mathrm{F}}^{2}\mathrm{d}t\Big/\int_{\mathbb{T}}\|\boldsymbol{\mathcal{B}}(t)\|_{\mathrm{F}}^{2}\,\mathrm{d}t=\mathcal{O}_{\mathrm{pr}}\big\{p_0^{-2\kappa}+n^{-1/2}p_0^{1/2}\big\}. \tag{3.26}$$

The first term decreasing with p_0 coincides with the traditional wisdom of nonparametric smoothing, while the second term increasing with p_0 reflects the complicated dimensionality in the tensor structure. It is suspected that the modeling of $\boldsymbol{\mathcal{B}}(\cdot)$ in (2.3) may need more smoothness, and a possible direction is considering $\boldsymbol{\mathcal{B}}(\cdot)$ to be a smooth map valued in a low-rank tensor space. We leave this as a potential future work.

In the following theorem, we demonstrate that the bound (3.25) is optimal in a minimax sense with respect to the tensor dimensionality. We do not tackle the functional aspect due to its complexity, which, as seen from (3.26), differs from the classical rate of convergence for nonparametric estimators.

Theorem 3. *Consider the model* $\mathfrak{M}$ *of* $(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta})$ *such that* $\mathscr{X}$ *satisfies* (3.17) *and* (3.19) *for* $\boldsymbol{\Upsilon} \in \bigcup_{\boldsymbol{s} \leq 2\boldsymbol{r}} \mathbb{M}_{\boldsymbol{s}}$ *and that* $\boldsymbol{\varepsilon} = \boldsymbol{y} - \mathscr{X}\boldsymbol{\Theta}$ *satisfies* $\|\boldsymbol{\varepsilon}\|^2 = \mathcal{O}_{\mathrm{pr}}(\xi)$ *with some* $\xi \geq \sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d$. *Then any estimator* $\widetilde{\boldsymbol{\Theta}}$ *for* $\boldsymbol{\Theta} \in \mathbb{M}_{\boldsymbol{r}}$ *based on* $\boldsymbol{y}$ *and* $\mathscr{X}$ *satisfies that*

$$\sup_{n \leq \prod_{d=0}^{D} p_d} n^{1/2} p_0^{-1/2} \sup_{(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}) \in \mathfrak{M}} \|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{\mathrm{F}} \geq 2^{-1/2} \|\boldsymbol{\mathcal{E}}_0\|_{\mathrm{F}}$$

for some random tensor $\boldsymbol{\mathcal{E}}_0$ *such that*

$$E(\|\boldsymbol{\mathcal{E}}_0\|_{\mathrm{F}}^2) \geq c_0 \Big(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d \Big)$$

where $c_0 > 0$ *is a constant depending only on* D.

As far as the difficulty of tensor computation is concerned, we end this section with the following remarks. Note that efficient tensor methods often turn multiplicative costs into additive.

Remark 1 (Initialization)**.** In Theorem 1, the quantity ζ_m in (3.21) is well defined since $\lim_{\boldsymbol{\Upsilon} \to \Theta} \mathscr{P}_{\boldsymbol{\Upsilon}} \boldsymbol{\Theta} = \boldsymbol{\Theta}$ (Luo and Zhang, 2023, Lemma 9). Theorem 1 and Corollary 1 require a suitable initialization, namely, $\|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}\|_{\mathrm{F}}$ is of order $\min\{\zeta_m, D^{-3/2} R_u^{-1/2} R_l^{1/2} \lambda_{\min}\}$. The linear dependence of the bound on $\lambda_{\min}$ matches the initialization condition in the literature (Luo and Zhang, 2023, Remark 2), and is often satisfied when warming up the output of some decomposition algorithms (Luo and Zhang, 2023, Section 4). In practice, we suggest setting $\boldsymbol{\Theta}^0 = \mathcal{H}_{\boldsymbol{r}}(\mathscr{X}^* \boldsymbol{y})$, which has worked well in our experiments.

Remark 2 (Sample size)**.** In view of (3.24), for estimation accuracy we need the sample size n to exceed $\mathcal{O}\{p_0^2(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d)C_m\}$. Note also that in order to fulfill (3.19), Rauhut et al. (2017, Theorem 2) suggests that $n = \mathcal{O}(\sum_{d=0}^{D} p_d r_d + \prod_{d=0}^{D} r_d)$ could be adequate if $\mathrm{rank}_{\mathrm{Tuc}}(\boldsymbol{\Upsilon}) = (r_0, r_1, \ldots, r_D)$ and the entries of $\boldsymbol{\Xi}_k$ are sub-Gaussian for $k = 1, \ldots, r_0$.

Remark 3 (Computational complexity)**.** In each iteration of the functional Riemannian Gauss–Newton algorithm, the computational cost for obtaining the update (2.16) is acceptable. Denote $\bar{p} = \max_d p_d$ and $\bar{r} = \max_d r_d$. Computing $\mathscr{R}^*_{\boldsymbol{\Theta}^k}\mathscr{Z}^*$ involves $\mathscr{R}^*_{\boldsymbol{\Theta}^k}\boldsymbol{\mathcal{Z}}_i$ for $i = 1, \ldots, n$, so the number of operations is $\mathcal{O}(n\bar{p}^{D+1}\bar{r})$. By Proposition 1, computing $\mathscr{R}^*_{\boldsymbol{\Theta}^k}\mathscr{A}\mathscr{R}_{\boldsymbol{\Theta}^k}$ only needs $\mathcal{O}(\bar{p}^2\bar{r})$ operations. Then it takes $\mathcal{O}[n\{\bar{r}^{D+1} + (D+1)\bar{p}\bar{r}\}^2]$ operations to solve the linear equation system in $\mathbb{D}$. Finally, extending the solution to $\mathrm{T}_{\boldsymbol{\Theta}^k}\mathbb{M}_{\boldsymbol{r}}$ uses $\mathcal{O}(\bar{p}^{D+1}\bar{r})$ operations. In summary, the cost is $\mathcal{O}(n\bar{p}^{D+1}\bar{r})$, provided that $\bar{r} = o(\bar{p}^{1/2})$.

4. Numerical Studies

We perform the proposed method on simulation and real data. In this section, the numerical results are reported.

4.1 Simulation

We generate a sample of $n = 500$ subjects with functional tensor covariates of order $(p_0, p_1, p_2) = (12, 8, 8)$. The functional tensors are $\boldsymbol{\mathcal{X}}_i(t) = \sum_{k=1}^{30} k^{-1}\sin(k\pi t)\boldsymbol{\Upsilon}_{ik}$, $i = 1, \ldots, n$, observed as $\boldsymbol{\mathcal{X}}_{ij}$ at the grid points $t_j = (j - 1/2)/p_0$, $j = 1, \ldots, p_0$. Here $\boldsymbol{\Upsilon}_{ik}$'s are i.i.d. random tensors with i.i.d. $\mathcal{N}(0,1)$ entries and the noise $\boldsymbol{\mathcal{E}}_{ij} = \boldsymbol{\mathcal{X}}_{ij} - \boldsymbol{\mathcal{X}}_i(t_j)$ are i.i.d. random tensors with i.i.d. $\mathcal{N}(0, 0.05^2)$ entries. The regression coefficient in (2.3) is generated by $\boldsymbol{\Theta} = \boldsymbol{\mathcal{S}} \times_{d=0}^{2} \boldsymbol{U}_d$ where $\boldsymbol{\mathcal{S}}$ is a random tensor of order $(r_0, r_1, r_2) = (2, 3, 3)$ with i.i.d. $\mathcal{N}(0,1)$ entries and $\boldsymbol{U}_d$'s are independent and uniformly distributed over $\mathbb{O}_{p_d,r_d}$, and cubic natural splines are used. Each response y_i is associated with an observational error ε_i from $\mathcal{N}(0, 0.1^2)$. The initialization for the functional Riemannian Gauss–Newton algorithm is chosen to be $\boldsymbol{\Theta}^0 = \mathcal{H}_{\boldsymbol{r}}(\mathscr{X}^*\boldsymbol{y})$. Given a tuning parameter ρ, we terminate the algorithm when the number of iterations reaches 80 or the relative error $\left\|\boldsymbol{\Theta}^k - \boldsymbol{\Theta}\right\|_{\mathrm{F}}/\|\boldsymbol{\Theta}\|_{\mathrm{F}}$ is less than 10^{-8}. The first panel of Figure 1 illustrates the quadratic convergence of our proposed algorithm in several Monte Carlo replications. Due to the observational error, the obtained estimates lie in a neighborhood of the true parameter. The adoption of a uniform grid could be relaxed, and we present an example of non-uniform grid points in the supplementary material, which achieves performance similar to the uniform case.

To select the tuning parameter ρ, we minimize the generalized cross-validation criterion,

$$\mathrm{GCV} = \frac{n^{-1}\|\boldsymbol{y} - \boldsymbol{H}_\rho \boldsymbol{y}\|^2}{\{1 - n^{-1}\operatorname{tr}(\boldsymbol{H}_\rho)\}^2},$$

where $\boldsymbol{H}_\rho = n^{-1}\mathscr{X}\mathscr{R}_{\widehat{\boldsymbol{\Theta}}}\left(n^{-1}\mathscr{R}^*_{\widehat{\boldsymbol{\Theta}}}\mathscr{X}^*\mathscr{X}\mathscr{R}_{\widehat{\boldsymbol{\Theta}}} + \rho\mathscr{R}^*_{\widehat{\boldsymbol{\Theta}}}\mathscr{A}\mathscr{R}_{\widehat{\boldsymbol{\Theta}}}\right)^{-1}\mathscr{R}^*_{\widehat{\boldsymbol{\Theta}}}\mathscr{X}^*$. The quality of our estimator is assessed by the relative integrated squared error,

$$\mathrm{RISE} = \int_0^1 \left\|\widehat{\boldsymbol{\mathcal{B}}}(t) - \boldsymbol{\mathcal{B}}(t)\right\|_{\mathrm{F}}^2 \mathrm{d}t \Big/ \int_0^1 \|\boldsymbol{\mathcal{B}}(t)\|_{\mathrm{F}}^2 \,\mathrm{d}t\,,$$

with $\widehat{\boldsymbol{\mathcal{B}}}(\cdot)$ being the plug-in estimator using the output $\widehat{\boldsymbol{\Theta}}$ of the algorithm. Under different choices of ρ, the results with correct Tucker rank are presented in the last two panels of Fig. 1. Note that $\rho = 0$ corresponds to tabular tensor regression. It can be seen that the introduction of a roughness penalty improves the estimation accuracy and the selection of the tuning parameter is valid. While the absolute difference of RISE appears small, this reflects a 1% relative improvement. For high-dimensional data, even marginal improvements in regularization efficacy can enhance interpretability by reducing overfitting. To the best of our knowledge, we propose the first method for functional tensor regression. Therefore, we compare the performance of our method with standard tensor regression methods.

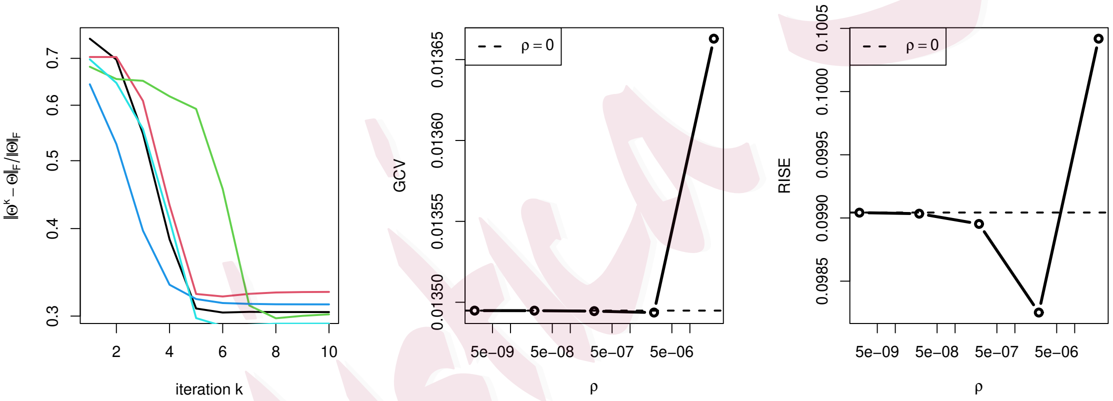


Figure 1: Left: Convergence performance of the functional Riemannian Gauss–Newton algorithm. Middle and Right: GCV and RISE versus the tuning parameter ρ. Displayed are averages based on 100 Monte Carlo replications of $(\boldsymbol{\mathcal{X}}_i, y_i)_{i=1,\ldots,500}$.

We then train the functional tensor regression model using different Tucker ranks. As shown in Table 1, the minimizer of GCV gives rise to correct specification of model rank and leads to fairly good estimation accuracy. When the number of rank candidates becomes excessively large, the shrinkage search method mentioned in Spencer et al. (2022) is beneficial for reducing computation costs.

Table 1: GCV and RISE corresponding to the selected tuning parameter under different Tucker ranks. Reported are the average and standard deviation (in the parenthesis) based on 100 Monte Carlo replications of $(\boldsymbol{\mathcal{X}}_i, y_i)_{i=1,\ldots,500}$. Marked with * is the true model rank.

Used rank	GCV × 10^2	RISE × 10^2	Used rank	GCV × 10^2	RISE × 10^2
(2,3,3)*	1.35(0.01)	9.83(0.28)	(5,5,5)	1.43(0.01)	73.67(1.40)
(2,4,3)	1.36(0.01)	10.88(0.36)	(2,3,2)	2.09(0.02)	30.48(0.40)
(3,3,3)	1.36(0.01)	16.40(0.43)	(2,4,2)	2.10(0.01)	30.90(0.43)
(3,4,3)	1.36(0.01)	20.80(0.49)	(3,3,2)	2.12(0.02)	37.11(0.59)
(3,4,4)	1.36(0.01)	26.61(0.57)	(3,4,2)	2.13(0.01)	42.25(0.73)
(4,4,4)	1.37(0.01)	37.68(0.83)	(2,2,2)	2.53(0.02)	44.27(0.43)
(4,4,3)	1.38(0.01)	27.96(0.59)	(3,2,2)	2.58(0.02)	50.25(0.49)

To assess the efficiency of the proposed method, we further conduct simulation studies with varying amplitude of observational errors: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. As σ ranges from 0.02 to 0.1, the Monte Carlo mean of $\frac{1}{n}\sum_{i=1}^{n}|y_i|$ varies from 0.348 to 0.358. The first panel of Figure 2 demonstrates that the RISE increases with the noise level in both functional and tabular tensor regression. However, the functional method consistently outperforms the tabular method across various signal-to-noise ratios. Since regularization methods usually present advantage by reducing variances, the proposed estimator shows marginally better robustness at higher noise levels.

The consistency of our estimator is illustrated in the second panel of Figure 2, where we revisit the $\sigma = 0.1$ scenario and incrementally adjust the sample size, documenting the corresponding changes in RISE. Notably, as expected, the RISE decreases, illustrating practical convergence.

Furthermore, we adjust the model based on the number of parameters. Specifically, with a fixed $n = 500$, we vary p_0 in the temporal mode from 3 to 18. The results, depicted in the third panel of Figure 2, show that initially, the RISE decreases due to effective regularization. Subsequently, as model complexity increases, RISE also increases. This provides partially a numerical verification for (3.26), which incorporates both functional and tensor aspects and cannot be accounted for using only a single perspective.

4.2 Real Data Example

Attention Deficit Hyperactivity Disorder (ADHD) is among the most common neurodevelopmental disorders of childhood. Zhou et al. (2013); Li et al. (2018) analyzed the ADHD data using tabular tensor regression models without considering the temporal effects. To remedy this, we apply our functional tensor regression model to the ADHD data. The original dataset can be downloaded from the ADHD-200 Sample Initiative (`http://fcon_1000.projects.nitrc.org/indi/adhd200/`), where the phenotypic test set consists of 197 subjects from 7 sites. Due to the compatibility of covariates, we extract the largest subsample from Peking University and

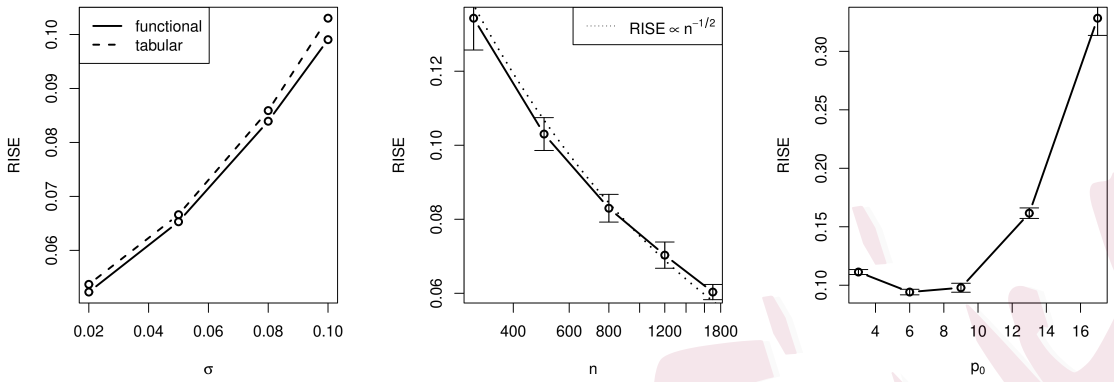


Figure 2: Left: RISE versus the standard deviation of errors. Displayed are averages based on 50 Monte Carlo replications. Solid and dashed lines correspond to functional and tabular tensor regression, respectively. Middle: RISE versus the sample size n. Solid line and error bars correspond to average and average ± standard deviation, respectively, all based on 50 Monte Carlo replications of $(\boldsymbol{\mathcal{X}}_i, y_i)_{i=1,\ldots,n}$. Dashed line reflects the theoretical result. The horizontal axis is in log scale to yield better visualization. Right: RISE versus the sampling frequency p_0. Displayed are average and average ± standard deviation based on 50 Monte Carlo replications.

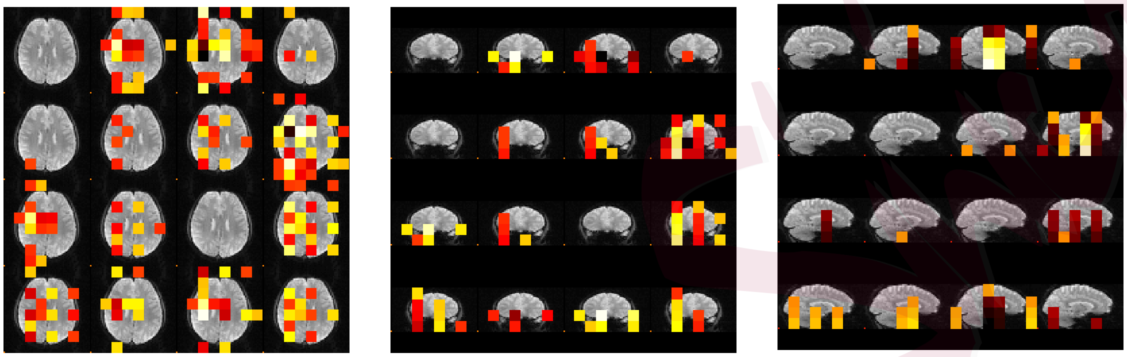
remove 1 subject whose ADHD index is missing. Then we obtain a sample of $n = 50$ subjects with ADHD index as the response. For each subject, there is an associated fMRI image that serves as the functional tensor covariate. To alleviate the computational burden, we pick $p_0 = 50$ consecutive time points, and reduce the spatial dimension to $p_1 \times p_2 \times p_3 = 8 \times 8 \times 4$ by assuming block-wise effects. Five scalar covariates are also included in the regression model: gender, age, medical status, verbal

IQ, and performance IQ.

We consider the functional tensor regression model

$$y = \boldsymbol{m}^\top \boldsymbol{\gamma} + \int_{\mathbb{T}} \langle \boldsymbol{\mathcal{X}}(t), \boldsymbol{\mathcal{B}}(t) \rangle \, \mathrm{d}t$$

where $\boldsymbol{m}$ denotes the vector of 1 and five scalar covariates. By profiling estimation using the ordinary least squares and the functional Riemannian Gauss–Newton algorithm, we fit the model with Tucker rank $(r_0, r_1, r_2, r_3) = (2, 2, 2, 2)$, which minimizes GCV and indicates the low-rank structure with 2 on each of the temporal direction and the three spatial directions. Figure 3 depicts the estimate of $\boldsymbol{\mathcal{B}}(\cdot)$ at a series of time points on the coronal, axial, and sagittal planes, respectively, along the temporal mode from left to right and from top to bottom, and reveals some regions with lasting significant influences on the ADHD index. We highlight the regions with point-in-time effects, with the more significant ones distinguished by a longer active time range. Two such regions are cortical surfaces and white matter, studied by Sowell et al. (2003); Makris et al. (2008). Cortical surfaces refer to the outer layer of the brain, known as the cerebral cortex, whose thickness and surface area are responsible for various cognitive functions including attention, memory, and executive functioning. White matter consists of myelinated nerve fibers that facilitate communication between different regions of the brain, and thus its abnormalities like changes in the integrity of white matter tracts could bring a disorder of brain network dysfunction. Although the mechanism of ADHD is not fully understood due to its complexity, our results imply a potential quantification. To make this more illustrative, the smooth time-varying effects of the indicated regions are plotted in Figure 4, in contrast to the cerebellum region that has little effect on ADHD index as the benchmark. The larger amplitudes corresponding to cortical surfaces and white matter reflect more important effects. Besides, the sinusoidal pattern arises from the interaction between the physiological cycle and the time-varying covariate effects.



Coronal Plane

Axial Plane

Sagittal Plane

Figure 3: Functional tensor regression applied to the ADHD data. Plotted are slices from three spatial dimensions where only coefficients with a magnitude larger than their 80% quantile are displayed. A brighter color means a larger value.

The variability of the ADHD data is well explained by our model, which can be seen from the R-squared. If $\bar{y}$ is the sample mean of y_i's and $\hat{y}_i$ denotes the fitted value of subject i, then

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 0.991.$$

To assess the out-of-sample prediction accuracy, we introduce the K-fold cross-validation,

$$\text{CV} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in S_k}(y_i - \hat{y}_i^{\backslash k})^2,$$

where the sample is split into K equal-sized parts $S_1, \ldots, S_K$ and $\hat{y}_i^{\backslash k}$ is the predicted value of subject i within the model trained by data from $S_1, \ldots, S_{k-1}, S_{k+1}, \ldots, S_K$.

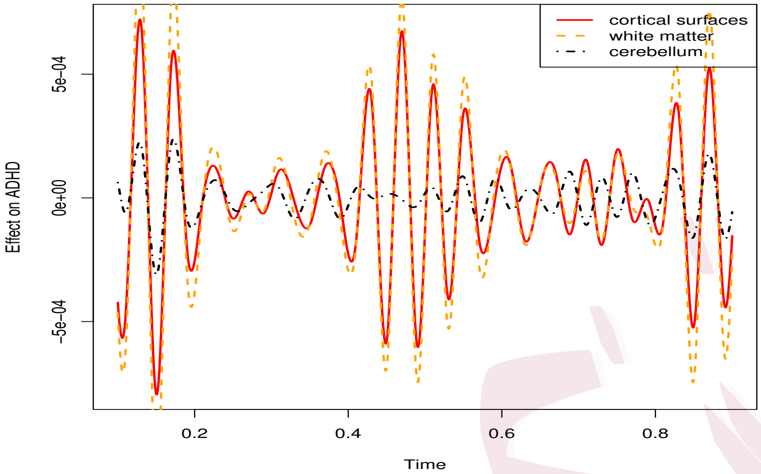


Figure 4: Estimated effects of different regions of the brain on ADHD along time.

Using $K = 10$, we have CV $= 121.9$ and CV $= 265.6$ for $\rho = 10^{2.3}$ (selected by GCV) and $\rho = 10^{-2}$, respectively. The tabular tensor regression corresponding to $\rho = 0$ is infeasible for computation due to the small sample size, so we use a sufficiently small $\rho = 10^{-2}$ to offer an approximation. This demonstrates that functional tensor regression improves upon the tabular method in the analysis of ADHD data.

A. T-HOSVD

Here we demonstrate the T-HOSVD operation (De Lathauwer et al., 2000) used in (2.16), as shown in Algorithm 2.

Algorithm 2: Truncated higher-order singular value decomposition

```
Input: tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{q_0 \times q_1 \times \cdots \times q_D}$, Tucker rank $\boldsymbol{r} = (r_0, r_1, \ldots, r_D)$
begin
    for $d = 0, 1, \cdots, D$ do
        Calculate $\tilde{\boldsymbol{U}}_d$ = leading $r_d$ columns of SVD$\{\mathscr{M}_d(\boldsymbol{\mathcal{T}})\}$;
    end
end
Output: $\mathcal{H}_{\boldsymbol{r}}(\boldsymbol{\mathcal{T}}) = \boldsymbol{\mathcal{T}} \times_{d=0}^{D} (\tilde{\boldsymbol{U}}_d \tilde{\boldsymbol{U}}_d^\top)$
```

B. List of Defined Constants

The following list indicates the location of the first occurrences of the constants we introduced.

- C_0: below (2.2)
- c_m: in (2.8)
- a, A: in Assumption 1
- C_φ: in Assumption 2
- R_u, R_l: in Lemma 1

Supplementary Materials

All proofs of the technical results and additional numerical results are collected in an online Supplementary Material. (.pdf file) The code and data are made available in a GitHub repository (`https://github.com/kellty/FTReg`).

Acknowledgements

Tongyu Li is the first author, and Fang Yao is the corresponding author. This research is partially supported by the National Key R&D Program of China (No. 2022YFA1003801), the National Natural Science Foundation of China (No. 12292981), the Newcorner Stone Science foundation through the Xplorer Prize, the LMAM and the Fundamental Research Funds for the Central Universities, Peking University (LMEQF).

References

Absil, P.-A., R. Mahony, and R. Sepulchre (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Ahmed, T., H. Raja, and W. U. Bajwa (2020). Tensor regression using low-rank and sparse tucker decompositions. *SIAM Journal on Mathematics of Data Science 2*(4), 944–966.

Bi, X., A. Qu, and X. Shen (2018). Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics 46*(6B), 3308–3333.

Bi, X., X. Tang, Y. Yuan, Y. Zhang, and A. Qu (2021). Tensors in statistics. *Annual review of statistics and its application 8*, 345–368.

Billio, M., R. Casarin, and M. Iacopini (2024). Bayesian markov-switching tensor regression for time-varying networks. *Journal of the American Statistical Association 119*(545), 109–121.

Billio, M., R. Casarin, M. Iacopini, and S. Kaufmann (2023). Bayesian dynamic tensor regression. *Journal* *of Business & Economic Statistics 41*(2), 429–439.

Boumal, N. (2023). *An introduction to optimization on smooth manifolds.* Cambridge University Press.

Cai, T. T. and P. Hall (2006). Prediction in functional linear regression. *The Annals of Statistics 34*(5), 2159–2179.

Chen, H., G. Raskutti, and M. Yuan (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research 20*(1), 172–208.

Chen, R., D. Yang, and C.-H. Zhang (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association 117*(537), 94–116.

Chen, Z., Y. Yang, and F. Yao (2024). Dynamic matrix recovery. *Journal of the American Statistical Association*, 1–12.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics 37*(1), 35–72.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications 21*(4), 1253–1278.

Durham, T. J., M. W. Libbrecht, J. J. Howbert, J. Bilmes, and W. S. Noble (2018). Predictd parallel epigenomics data imputation with cloud-based tensor decomposition. *Nature communications 9*(1), 1402.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing.* CRC press.

Gandy, S., B. Recht, and I. Yamada (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse problems 27*(2), 025010.

Goldfarb, D. and Z. Qin (2014). Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal* *on Matrix Analysis and Applications 35*(1), 225–253.

Guhaniyogi, R. (2014). Bayesian methods for tensor regression. *Wiley StatsRef: Statistics Reference Online*, 1–18.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). Bayesian tensor regression. *Journal of Machine Learning Research 18*(79), 1–31.

Hackbusch, W. (2019). *Tensor spaces and numerical tensor calculus* (2 ed.), Volume 56. Springer.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Han, R., P. Shi, and A. R. Zhang (2024). Guaranteed functional tensor singular value decomposition. *Journal of the American Statistical Association 119*(546), 995–1007.

Han, R., R. Willett, and A. R. Zhang (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics 50*(1), 1–29.

Han, Y., D. Yang, C.-H. Zhang, and R. Chen (2024). Cp factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae036.

Hillar, C. J. and L.-H. Lim (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM) 60*(6), 1–39.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The annals of applied statistics 9*(3), 1169.

Jing, B.-Y., T. Li, Z. Lyu, and D. Xia (2021). Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics 49*(6), 3181–3205.

Kim, T.-K., S.-F. Wong, and R. Cipolla (2007). Tensor canonical correlation analysis for action classification.

In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE.

Koch, O. and C. Lubich (2010). Dynamical tensor approximation. *SIAM Journal on Matrix Analysis and Applications 31*(5), 2360–2375.

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

Li, L. and X. Zhang (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association 112*(519), 1131–1146.

Li, X., D. Xu, H. Zhou, and L. Li (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences 10*(3), 520–545.

Liu, J., C. Zhu, Z. Long, Y. Liu, et al. (2021). Tensor regression. *Foundations and Trends® in Machine Learning 14*(4), 379–565.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics 27*(3), 638–647.

Lu, C., J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan (2019). Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE transactions on pattern analysis and machine intelligence 42*(4), 925–938.

Lu, H., K. N. Plataniotis, and A. N. Venetsanopoulos (2008). Mpca: Multilinear principal component analysis of tensor objects. *IEEE transactions on Neural Networks 19*(1), 18–39.

Luo, Y., D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen (2015). Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE transactions on Knowledge and Data Engineering 27*(11), 3111–3124.

Luo, Y. and A. R. Zhang (2023). Low-rank tensor estimation via riemannian gauss-newton: Statistical optimality and second-order convergence. *Journal of Machine Learning Research 24*(381), 1–48.

Makantasis, K., A. D. Doulamis, N. D. Doulamis, and A. Nikitakis (2018). Tensor-based classification models for hyperspectral data analysis. *IEEE Transactions on Geoscience and Remote Sensing 56*(12), 6884–6898.

Makris, N., S. L. Buka, J. Biederman, G. M. Papadimitriou, S. M. Hodge, E. M. Valera, A. B. Brown, G. Bush, M. C. Monuteaux, V. S. Caviness, et al. (2008). Attention and executive systems abnormalities in adults with childhood adhd: a dt-mri study of connections. *Cerebral cortex 18*(5), 1210–1220.

Phan, A. H. and A. Cichocki (2010). Tensor decompositions for feature extraction and classification of high dimensional datasets. *Nonlinear theory and its applications, IEICE 1*(1), 37–68.

Ramsay, J. and B. Silverman (2005). *Functional data analysis* (2 ed.). Springer.

Ramsay, J. O. and C. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(3), 539–561.

Rauhut, H., R. Schneider, and Ž. Stojanac (2017). Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications 523*, 220–262.

Sowell, E. R., P. M. Thompson, S. E. Welcome, A. L. Henkenius, A. W. Toga, and B. S. Peterson (2003). Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder. *The Lancet 362*(9397), 1699–1707.

Spencer, D., R. Guhaniyogi, R. Shinohara, and R. Prado (2022). Bayesian tensor regression using the tucker decomposition for sparse spatial modeling. *arXiv preprint arXiv:2203.04733*.

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

Sun, W. W. and L. Li (2019). Dynamic tensor clustering. *Journal of the American Statistical Association 114*(528), 1894–1907.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika 31*(3), 279–311.

Uschmajew, A. and B. Vandereycken (2013). The geometry of algorithms using hierarchical tensors. *Linear Algebra and its Applications 439*(1), 133–166.

Utreras, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numerische Mathematik 42*(1), 107–117.

Vannieuwenhoven, N., R. Vandebril, and K. Meerbergen (2012). A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing 34*(2), A1027–A1052.

Zhang, C., R. Han, A. R. Zhang, and P. M. Voyles (2020). Denoising atomic resolution 4d scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy 219*, 113123.

Zhang, Y., X. Bi, N. Tang, and A. Qu (2021). Dynamic tensor recommender systems. *The Journal of Machine Learning Research 22*(1), 3032–3066.

Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(2), 463–483.

Zhou, H., L. Li, and H. Zhu (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association 108*(502), 540–552.

Zhou, J., W. W. Sun, J. Zhang, and L. Li (2023). Partially observed dynamic tensor response regression. *Journal of the American Statistical Association 118*(541), 424–439.

Zhou, Y. and H.-G. Müller (2022). Network regression with graph laplacians. *Journal of Machine Learning Research 23*(320), 1–41.

Zhou, Y., R. K. Wong, and K. He (2021). Tensor linear regression: Degeneracy and solution. *IEEE Access 9*, 7775–7788.

Zhou, Y., R. K. Wong, and K. He (2024). Broadcasted nonparametric tensor regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae027.

School of Mathematical Sciences, Center for Statistical Science, Peking University, China

E-mail: tongyuli@pku.edu.cn

School of Mathematical Sciences, Center for Statistical Science, Peking University, China

E-mail: fyao@math.pku.edu.cn

Department of Biostatistics & Bioinformatics and Department of Computer Science, Duke University, U.S.A.

E-mail: anru.zhang@duke.edu