BROADCASTED NONPARAMETRIC TENSOR REGRESSION

A Thesis

by

YA ZHOU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Raymond Ka Wai Wong
Committee Members,	Xianyang Zhang
	Yang Ni
	James Caverlee
Head of Department,	Jianhua Z. Huang

December 2019

Major Subject: Statistics

ABSTRACT

Uncovering relationships among different variables from tensor data often lead to enhanced understanding of scientific and engineering problems. One recent statistical development under this setup is tensor regression. Most of the works make a strong assumption that the tensor covariates enter the model linearly, which is rather restrictive. Those models that consider the nonlinearity suffer from the curse of dimensionality and possess very weak interpretability. Motivated by observations from many real life applications and the need for nonlinearity, we propose a nonparametric tensor regression with broadcasting structure. Within the proposed model framework, we develop both an alternating updating algorithm as well as the asymptotic convergence rate for the proposed estimation. Through experiments on the synthetic data and two real data, we demonstrate the power of the proposed broadcasted nonparametric tensor regression.

DEDICATION

To my parents and girlfriend, for their endless love, support and encouragement.

ACKNOWLEDGMENTS

I want to thank my advisor, Professor Raymond Ka Wai Wong. He is a fantastic advisor, mentor, and friend. He supervised me with his patience and deep understanding of statistical theory. I've learned a lot from his high-level insights in science and his passion for developing interpretable and practically useful methodologies. I would also like to thank Professor Kejun He, for much useful advice on research, paper, presentation, and many other aspects. He told me his experience living in College Station and cares me both academically and personally.

I'm very grateful to my committee members, Professors Xianyang Zhang, Yang Ni, and James Caverlee, for their willingness to serve on my committee. Thank them for their valuable comments and suggestions provided during my oral defense.

One special thank should go to my parents and girlfriend for their endless love, support, and encouragement.

Lastly, I want to thank all my friends, classmates, and teachers who have ever helped me.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professors Raymond Ka Wai Wong (advisor), Xianyang Zhang and Yang Ni of the Department of Statistics and Professor James Caverlee of the Department of Computer Science and Engineering.

This thesis is completed by the student under the direction of Professor Raymond Ka Wai Wong. Part of the work is a collaborative product with Professor Raymond Ka Wai Wong and Professor Kejun He.

Funding Sources

There are no outside funding for the research.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Nowadays, tensor data are abundant in many different areas, such as clinical applications (Wang et al., 2014), computer vision (Lu et al., 2013), genomics (Durham et al., 2018), neuroscience (Zhou et al., 2013) and recommender systems (Zhu et al., 2018). Uncovering relationships among different variables from tensor data often lead to enhanced understanding of scientific and engineering problems. One recent statistical development under this setup is tensor regression. In this work, we focus on models that involve a tensor covariate $\mathbf{X} = (X_{i_1,i_2,\ldots,i_D}) \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ of order D. Notice that tensor regression based on vector covariate (e.g., Sun and Li, 2017; Li and Zhang, 2017; Hu et al., 2019) is also a popular research direction.

In the literature, there are roughly three categories of tensor regression with tensor covariates according to the response type. The first is scale-on-tensor regression, where the response is a scalar (e.g., Zhou et al., 2013; Zhao et al., 2014; Hou et al., 2015; Chen et al., 2019). Within this category, there are methods that focus particularly on image covariates (e.g., Reiss and Ogden, 2010; Zhou and Li, 2014; Wang et al., 2017; Kang et al., 2018). The second is vector-on-tensor regression, in which one of the response is a vector (e.g., Miranda et al., 2018). The last one is tensor-on-tensor regression, with a tensor output (e.g., Hoff, 2015; Lock, 2018; Raskutti et al., 2019).

The majority of the above models make a strong assumption that the tensor covariates enter the model linearly (or, for non-Guassian response, via a known link function as in generalized linear models). To date, very few works go beyond linearity. According to the above categorization, they all fall into the first category of tensor regression. On the application side, Zhao et al. (2013) and Hou et al. (2015) used a Gaussian process regression model to catch possible nonlinear effects of tensor covariates for better prediction in video surveillance applications and neuroimaging analyses. Their approaches rely on the choice of kernel function defined on tensors. One could flatten a tensor into a high-dimensional vector and adopt popular kernels on vectors, such as Gaussian kernel. However, this would ignore the structural information of the tensor and also suffer from the

curse of dimensionality. Zhao et al. (2014) proposed the use of a kernel based on matricizations of the tensor covariates. But the corresponding discussion is brief, and no theoretical justification is presented in their work. More importantly, their method lacks interpretable and efficient parametrization, such as low-rank representations. That said, these early efforts demonstrate the power and the need for nonparametric modeling in various applications.

Another class of methods incorporates nonlinearity through a more explicit function space by imposing low-rank structures on covariates. Kanagawa et al. (2016) proposed a regression model for a rank-1 tensor covariate, i.e., $\mathbf{X} = \mathbf{x}_1 \circ \mathbf{x}_2 \circ \cdots \circ \mathbf{x}_D$, where $\circ$ represents an outer product. Imaizumi and Hayashi (2016) extended their work to higher-rank tensor covariates and proposed the model

$$m(\mathbf{X}) = \sum_{r=1}^{R} \sum_{q=1}^{Q} \lambda_q \prod_{d=1}^{D} g_{d,r}(\mathbf{x}_{q,d}), \tag{1.1}$$

where $\mathbf{X}$ is assumed to have a smallest CANDECOMP/PARAFAC (CP) decomposition

$$\mathbf{X} = \sum_{q=1}^{Q} \lambda_q \mathbf{x}_{q,1} \circ \mathbf{x}_{q,2} \circ \cdots \circ \mathbf{x}_{q,D}, \tag{1.2}$$

where $\|\mathbf{x}_{q,d}\|_2 = 1$ and $\lambda_Q \geq \lambda_{Q-1} \geq \cdots \geq \lambda_1 \geq 0$. When $Q = 1$, (1.1) recovers the model of Kanagawa et al. (2016). Due to difficulty in estimation, a small Q should be used. However, in most cases, the tensor covariate is not exactly low-rank, and the rank of the covariate usually varies from observation to observation. Although the additive form of (1.1) has significantly reduced model complexity, the function $g_{d,r}$ is still an intrinsically $(p_d - 1)$-dimensional function, which may still be difficult to estimate. For example, given a $64 \times 64 \times 64$ 3D-image covariate, $p_1 = p_2 = p_3 = 64$. This also aligns with a finding (Imaizumi and Hayashi, 2016) that the asymptotic convergence rate grows exponentially with $\max_d p_d$. Furthermore, the model is difficult to interpret due to its dependence on the CP representation of the covariate, which may be non-unique (Stegeman and Sidiropoulos, 2007).

Overall, although the above nonlinear models demonstrate successes in certain applications, they still suffer from the curse of dimensionality and possesses very weak interpretability. In this

work, we propose an alternative that addresses both of these issues.

Our proposed model extends the low-rank tensor linear model developed by Zhou et al. (2013), which we briefly describe as follows. Given a vector covariate $\mathbf{z} \in \mathbb{R}^{p_0}$, a tensor covariate $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ and a response variable $y \in \mathcal{Y} \subseteq \mathbb{R}$. Zhou et al. (2013) proposed a generalized linear model with the following form of linear predictor

$$g\{\mathbb{E}(y|\mathbf{z}, \mathbf{X})\} = \nu + \boldsymbol{\gamma}^\intercal \mathbf{z} + \langle \mathbf{B}, \mathbf{X} \rangle,$$

where g is a link function and, $\nu \in \mathbb{R}$, $\gamma \in \mathbb{R}^{p_0}$ and $\mathbf{B} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ are parameters. In particular, the coefficient tensor $\mathbf{B}$ is assumed to admit a CP decomposition

$$\mathbf{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D},$$

where $\boldsymbol{\beta}_{r,d} \in \mathbb{R}^{p_d}$ and R is the rank. Combined with sparsity-inducing regularization, Zhou et al. (2013) and Zhou and Li (2014) showed that low-rank coefficient tensor $\mathbf{B}$ can be used to infer the region (entries) of $\mathbf{X}$ that explains the response.

In this work, to formulate a nonparametric regression technique that accommodates tensor predictors, we propose a nonparametric tensor regression with broadcasting structure (to be defined below). In many real life applications, entries within some regions of the tensor (especially images) share similar effects due to certain spatial structures such as a spatially clustered effect. For instance, Zhou et al. (2013) showed that voxels within two brain sub-regions have similar linkages with attention deficit hyperactivity disorder. Miranda et al. (2018) demonstrated that voxels within several sub-regions of the brain have a spatially clustered effect on Alzheimer's disease. Motivated by these observations and the need of nonlinearity, we propose to "broadcast" similar nonlinear relationship (with the response) to different entries of the tensor covariate. On a high-level, we model the nonlinearity effect by uni-dimensional nonparametric functions, which are supposed to be functions applied to an individual entry. These uni-dimensional functions are then shared by every entry. We call this operation of distributing a uni-dimensional function to all entries "broad-

casting". Additional scaling coefficients are used to linearly scale the effect of the uni-dimensional functions. Through regularizing these scaling coefficients, we can restrict the effects of certain uni-dimensional functions to smaller regions. As shown by Zhou et al. (2013) and Zhou and Li (2014), lasso-type regularization alone may result in poor performance in region selection, while an additional low-rank constraint/regularization would produce more successful results. Therefore we also restrict the scaling coefficient to be low-rank.

Within the proposed model framework, we develop both an alternating updating algorithm as well as the asymptotic convergence rate for the proposed estimation. Our theory includes tensor linear model (Zhou et al., 2013) as a special case. However, unlike Zhou et al. (2013), ours is of high-dimensional nature, which allows $p_1, \ldots, p_D$ to diverge. We believe this asymptotic framework is more relevant to many applications where the data (e.g., imaging data) involves large p_j's when compared to the sample size. Through two real data examples, we demonstrate the power of the proposed broadcasted nonparametric tensor regression. Overall, the proposed method timely responds to a number of growing needs of catching nonlinearity with interpretable models and rigorous theoretical developments.

The rest of the article is organized as follows. Section 2 introduces the broadcasted nonparametric model. The proposed estimation and computational algorithm, and corresponding theoretical results are presented in Sections 3 and 4. The practical performance of the proposed method is illustrated via both a simulation study and two real data applications, all presented in Section 5. Technical proofs are delegated to Appendix A.

2. MODEL

Consider $\mathbf{X} \in \mathcal{X} = \prod_{i_1,i_2,\ldots,i_D=1}^{p_1,\ldots,p_D} \mathcal{X}_{i_1,\ldots,i_D}$ where $\mathcal{X}_{i_1,\ldots,i_D}$ is a compact subset of $\mathbb{R}$. Without loss of generality, we assume $\mathcal{X} = [0,1]^{p_1 \times p_2 \times \cdots \times p_D}$. For simplicity, we focus on the additive error model

$$y = m(\mathbf{X}) + \epsilon, \tag{2.1}$$

where $m : \mathcal{X} \to \mathbb{R}$ is an unknown regression function of interest and ϵ is a random error of mean zero. The observed data $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$ are modeled as i.i.d. copies of $(y, \mathbf{X})$. Our first task is to propose a useful and interpretable model for the regression function m.

2.1 Common nonparametric strategies: curse of dimensionality

As discussed in Section 1, existing works of nonparametric tensor regression suffer from the curse of dimensionality and lack good interpretability. Here we briefly discuss several common nonparametric regression models for vector covariates. A direct application of these models relies on flattening the tensor into a vector, which non-ideally ignores the tensor structure.

Let us begin with the most general model in which $m(\cdot)$ is an unstructured (smooth) function mapping $\mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ to $\mathbb{R}$. Despite the flexibility, this model unsurprisingly suffers heavily from the curse of dimensionality. For a typical $64 \times 64 \times 64$ image, we are facing a nonparametric estimation of a function with dimension 64^3, which is generally impractical.

A common alternative in the literature of nonparametric regression is to assume an additive form (e.g., Stone, 1985; Hastie and Tibshirani, 1990; Wood, 2006):

$$m(\mathbf{X}) = \frac{1}{s} \sum_{i_1,i_2,\cdots,i_D} m_{i_1 i_2 \cdots i_D}(X_{i_1 i_2 \cdots i_D}),$$

where $s = \prod_{d=1}^D p_d$ is the number of entries in the tensor. This model however involves s (e.g., 64^3 in the above example) uni-dimensional functions. Potentially, sparsity (e.g., Lin et al., 2006; Meier et al., 2009; Ravikumar et al., 2009; Huang et al., 2010; Raskutti et al., 2012; Fan et al.,

2011; Chen et al., 2018) could be introduced to help. But typical sparse estimations, when applied to a tensor covariate, would ignore important tensor structures and may allow only too few pixels to have effect, especially when the sample size n is much smaller than s.

Another class of common models is the single index model (e.g., Ichimura, 1993; Horowitz and Härdle, 1996):

$$m(\mathbf{X}) = f\left(\sum_{i_1,\ldots,i_D} a_{i_1,i_2,\ldots,i_D} X_{i_1,i_2,\ldots,i_D}\right),$$

where f is an unknown uni-dimensional function and $\{a_{i_1,i_2,\ldots,i_D}\}$ are s unknown weight parameters. Although there is only one uni-dimensional function, this model involves abundant coefficient parameters, often much more than the sample size. One could impose sparsity to the coefficients (e.g. Alquier and Biau, 2013; Radchenko, 2015). However, similar issues of ignoring tensor structures, as in sparse additive model, also occur here. These problems would be worsened in more complicated index models such as additive index model and multiple index model.

In the following subsection, we propose a novel and economical model which makes use of the tensor structure. Our model has a close relationship with the additive models, but do not suffer from the above problems of the additive models.

2.2 Low-rank modeling with broadcasting

As mentioned above, the additive models involve too many functions. A simple remedy is to restrict all entries to share the same function: $m(\mathbf{X}) = s^{-1}\sum_{i_1,i_2,\cdots,i_D} f(X_{i_1 i_2 \cdots i_D})$. In other words, we *broadcast*[1] the same function f to every entry. In many real life applications, entries within some regions of the tensor (especially images) share similar effects due to certain spatial structures such as a spatially clustered effect. For instance, (Zhou et al., 2013) showed that voxels within two brain sub-regions have similar linkages with attention deficit hyperactivity disorder. (Miranda et al., 2018) demonstrated that voxels within several sub-regions of the brain have a spatially clustered effect on Alzheimer's disease. Hence, broadcasting a nonlinear relationship (with the response) is a well-motivated modeling strategy. But the assumption that *every* entry has

[1] A term widely used for similar operations in programming languages such as Python.

the same (nonlinear) effect on the response is very restrictive. Specifically, in many imaging data, there are usually only one or a few clusters of entries that are related to the response. Moreover, these regions may have different nonlinear effects to the response.

For any two tensors $\mathbf{A} = (A_{i_1,\ldots,i_D}), \mathbf{B} = (B_{i_1,\ldots,i_D})$ of the same dimensions, we define $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i_1,\ldots,i_D} A_{i_1,\ldots,i_D} B_{i_1,\ldots,i_D}$. Motivated by Zhou et al. (2013), we utilize the (low-rank) tensor structure to discover important regions of the tensor so as to broadcast a nonparametric modeling on such regions. We propose the following broadcasted nonparametric regression model:

$$m(\mathbf{X}) = \nu + \frac{1}{s} \sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle \,, \tag{2.2}$$

where $\nu \in \mathbb{R}$, $\boldsymbol{\beta}_{r,d} \in \mathbb{R}^{p_d}$, and $F_r : \mathbb{R}^{p_1 \times \cdots \times p_D} \to \mathbb{R}^{p_1 \times \cdots \times p_D}$ is defined by broadcasting:

$$(F_r(\mathbf{X}))_{i_1 i_2 \cdots i_D} = f_r(\mathbf{X}_{i_1 i_2 \cdots i_D}) \quad \text{with } f_r : \mathbb{R} \to \mathbb{R},$$

i.e., the $(i_1, \ldots, i_D)$-th entry of $F_r(\mathbf{X})$ is $f_r(X_{i_1 i_2 \cdots i_D})$. Here $f_r \in \mathcal{H}$ admits a nonparametric modeling specified by the (infinite-diemnsional) function class $\mathcal{H}$. Following the convention (e.g., Stone, 1985), $\mathcal{H}$ is assumed to be a smooth function class with some Hölder condition with details specified in Section 4. In this model, there are R different components, each of which is composed of a uni-dimensional function f_r to be broadcasted, and a rank-1 scaling (coefficient) tensor $\boldsymbol{\beta}_{r,1} \circ \cdots \circ \boldsymbol{\beta}_{r,D}$ to linearly scale the effect across different entries.

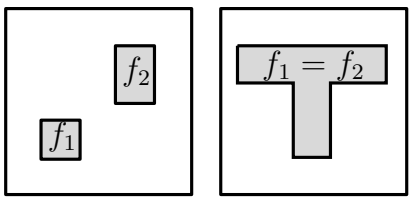

Figure 2.1: Examples of the broadcasted model

The model is economical since these broadcasted functions are uni-dimensional and these scal-

ing tensors are of rank 1. Several components can be combined to characterize different nonlinear effects adapted to different subregions (if appropriate sparse estimation on scaling tensors is imposed). We give two simple examples of $D = 2$ depicted in Figure 2.1, where the shaded regions correspond to nonzero entries in corresponding scaling tensors. In the left figure, there are two rank-1 regions (shaded) with different nonlinear functions; in the right figure, there is a rank-2 region formed by two scaling tensors with a shared nonlinear effect ($f_1 = f_2$).

Similar to the tensor linear model (Zhou et al., 2013), the proposed model suffers from parameter identifiability issues, i.e., broadcasted functions and scaling tensors. For instance, one can multiply $\boldsymbol{\beta}_{r,1}$ by 10, and divide $\boldsymbol{\beta}_{r,2}$ by 10, but still obtain the same $m(\cdot)$. Another example is a permutation of the components. To understand the nonlinear effect of entries, only the identification of $m(\cdot)$ is needed and thus such non-identifiability is in general, not an issue. For completeness, we provide sufficient conditions for the parameter identification, similar to the Kruskal's condition (Kruskal, 1989; Sidiropoulos and Bro, 2000). As the discussion is lengthy and not directly related to the following sections (where only identification of $m(\cdot)$ is needed), we refer interested readers to Appendix A.1.

3.1 Spline approximation and penalized estimation

The broadcasted functions f_r, $r = 1, \ldots, R$, will be approximated by B-spline functions of order ζ, i.e.,

$$f_r(x) \approx \sum_{k=1}^{K} \alpha_{r,k} b_k(x), \tag{3.1}$$

where $\mathbf{b}(x) = (b_1(x), \cdots, b_K(x))^\intercal$ is a B-spline basis and $\alpha_{r,k}$'s are the corresponding spline coefficients. By writing $\boldsymbol{\alpha}_r = (\alpha_{r,1}, \ldots, \alpha_{r,K})^\intercal$ and ignoring the spline approximation error, the regression function (2.2) can be written as

$$m(\mathbf{X}) = \nu + \frac{1}{s} \sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \boldsymbol{\alpha}_r, \Phi(\mathbf{X}) \rangle \,, \tag{3.2}$$

where $\Phi(\mathbf{X})$ is $p_1 \times p_2 \times \cdots \times p_D \times K$-dimensional tensor function such that its $(i_1, \ldots, i_D, k)$'s entry satisfies $(\Phi(\mathbf{X}))_{i_1,\ldots,i_D,k} = b_k(X_{i_1 \cdots i_D})$. In accordance with the model identifiability conditions $\int_0^1 f_r(x)\mathrm{d}x = 0$, $r = 1, \ldots, R$, the coefficients of basis functions are subject to

$$\int_0^1 \sum_{k=1}^{K} \alpha_{r,k} b_k(x)\mathrm{d}x = 0, \quad r = 1, \ldots, R.$$

Letting $u_k = \int_0^1 b_k(x)\mathrm{d}x$, the objective function $m(\mathbf{X})$ can be estimated through solving

$$\begin{aligned} &\arg\min_{\nu, \mathbf{A}} \sum_{i=1}^{n} \left(y_i - \nu - \frac{1}{s} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle \right)^2 \\ &\text{s.t.} \quad \mathbf{A} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \boldsymbol{\alpha}_r \\ &\qquad \sum_{k=1}^{K} \alpha_{r,k} u_k = 0, \quad r = 1, \ldots, R. \end{aligned} \tag{3.3}$$

Directly solving (3.3) is not computationally efficient since it involves too many linear con-

straints. To further simplify the optimization problem, we propose to remove the constraints by using the equivalent truncated power basis (Ruppert et al., 2003). We let $\{b'_k(x)\}_{k=1}^K$ denote the truncated power basis with the same order and interior knots as $\{b_k(x)\}_{k=1}^K$, where $b'_1(x)$ is the constant function. Theorem 4 in Appendix yields that the constrained optimization problem (3.3) is equivalent to

$$\begin{aligned} &\arg\min_{\tilde{\nu},\tilde{\mathbf{A}}} \sum_{i=1}^{n} \left(y_i - \tilde{\nu} - \frac{1}{s}\langle \tilde{\mathbf{A}}, \tilde{\Phi}(\mathbf{X}_i)\rangle \right)^2 \\ &\text{s.t.} \quad \tilde{\mathbf{A}} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \tilde{\boldsymbol{\alpha}}_r, \end{aligned} \tag{3.4}$$

where $\tilde{\Phi}(\mathbf{X}) \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K-1}$ with $(\tilde{\Phi}(\mathbf{X}))_{i_1,\ldots,i_D,k} = b'_{k+1}(X_{i_1 \cdots i_D})$, $k = 1, \ldots, K-1$, and $\tilde{\boldsymbol{\alpha}}_r$ is the vector of coefficients. In other words, the constraints are removed by reducing one degree freedom of the basis functions.

Although the low rank structure can help identify the important region (Zhou et al., 2013; Zhou and Li, 2014), we propose to add an additional regularization term to improve the performance of the estimation, especially when sample size is relatively small. In particular, we consider the following penalized estimation

$$\begin{aligned} &\arg\min_{\tilde{\nu},\tilde{\mathbf{A}}} \sum_{i=1}^{n} \left(y_i - \tilde{\nu} - \frac{1}{s}\langle \tilde{\mathbf{A}}, \tilde{\Phi}(\mathbf{X}_i)\rangle \right)^2 + \sum_{r=1}^{R} \sum_{d=1}^{D} \sum_{i=1}^{p_d} P_\lambda(\beta_{r,di}) \\ &\quad \text{s.t.} \quad \tilde{\mathbf{A}} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \tilde{\boldsymbol{\alpha}}_r \\ &\qquad\quad \|\tilde{\boldsymbol{\alpha}}_r\|_2^2 = 1, \quad r = 1, \ldots, R, \end{aligned} \tag{3.5}$$

where $P_\lambda(\cdot)$ is the penalty function with penalized parameter λ. Typical choices of the penalty function in the scope of linear regression include the Lasso penalty (Tibshirani, 1996), the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), the elastic net penalty (Zou and Hastie, 2005), and minimax concave penalty (MCP) (Zhang, 2010). Among them, the elastic net penalty can identify the relevant predictors as well as the Lasso penalty in the case $p \gg n$, but also deliver good prediction performance when the number of predictors are moderate and the variables

are highly correlated, which usually happen in neuroimaging data (Zhou and Li, 2014). In other words, we consider

$$P_\lambda(\beta_{r,di}) = \lambda_1 \left\{ \frac{1}{2}(1-\lambda_2)\beta_{r,di}^2 + \lambda_2 |\beta_{r,di}| \right\},$$

where $\lambda_2 \in [0,1]$ and $\lambda_1 \in \mathbb{R}^+$. In (3.5), the norm 1 restrictions for $\tilde{\boldsymbol{\alpha}}_r$'s are used to regularize the coefficients of truncated power basis.

3.2 Computational algorithm

We propose to use a scale-adjusted block-wise descent algorithm to solve (3.5) as follows. Recall $\mathbf{B}_d = (\boldsymbol{\beta}_{1,d}, \ldots, \boldsymbol{\beta}_{R,d})$, $d = 1, \ldots, D$. Analogously, we denote $\tilde{\mathbf{B}}_{D+1} = (\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_R)$. For convenience, we let

$$\boldsymbol{\theta} = (\tilde{\nu}, \mathbf{B}_1, \ldots, \mathbf{B}_D, \tilde{\mathbf{B}}_{D+1}),$$

and denote the least squares term, the penalty term, and the whole objective function as

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - \tilde{\nu} - \frac{1}{s} \sum_{r=1}^{R} \left\langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \tilde{\boldsymbol{\alpha}}_r, \tilde{\Phi}(\mathbf{X}_i) \right\rangle \right)^2,$$

$$G(\boldsymbol{\theta}) = \sum_{r=1}^{R} \sum_{d=1}^{D} \sum_{i=1}^{p_d} P_\lambda(\beta_{r,di}),$$

and $LG(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + G(\boldsymbol{\theta})$, respectively. Observe that

$$\begin{aligned} \sum_{r=1}^{R} \left\langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \tilde{\boldsymbol{\alpha}}_r, \tilde{\Phi}(\mathbf{X}) \right\rangle &= \left\langle \mathbf{B}_d, \tilde{\Phi}(\mathbf{X})_{(d)} \mathbf{B}_{-d} \right\rangle \\ &= \left\langle \mathrm{vec}\{\tilde{\Phi}(\mathbf{X})_{(d)} \mathbf{B}_{-d}\}, \mathrm{vec}(\mathbf{B}_d) \right\rangle, \end{aligned}$$

where $\mathbf{B}_{-d} = \mathbf{B}_1 \odot \cdots \odot \mathbf{B}_{d-1} \odot \mathbf{B}_{d+1} \odot \cdots \odot \mathbf{B}_{D+1}$ and $\tilde{\Phi}(\mathbf{X})_{(d)}$ is the mode-d matricization (Kolda and Bader, 2009) of tensor $\tilde{\Phi}(\mathbf{X})$. We can thus alternatively update $\mathbf{B}_d$, $d = 1, \cdots, D$, by the elastic net penalized linear regression (Zou and Hastie, 2005). As for $\mathbf{B}_{D+1}$, when we use the norm-homogeneous penalty, such as the elastic net, it can be relaxed to a standard quadratically constrained quadratic program (QCQP, Boyd and Vandenberghe, 2004). Therefore, the dual ascent

method (Boyd et al., 2011) and second-order cone programming (Alizadeh and Goldfarb, 2003) can be used for the block-wise updating.

One manipulation of the least squares term $L(\boldsymbol{\theta})$ is the scale shift among $\boldsymbol{\beta}_{r,d}$'s for $d = 1, \cdots, D$, i.e., the scale of $\boldsymbol{\beta}_{r,d_1}$ can shift to $\boldsymbol{\beta}_{r,d_2}$, $d_1 \neq d_2$, without changing the value of the least squares term. This manipulation can, however, change the value of penalty term $G(\boldsymbol{\theta})$. We propose an optimal rescaling strategy for the elastic net penalty. Specifically, we assume $\boldsymbol{\beta}_{r,d} \neq \mathbf{0}$, $r = 1, \ldots, R$, $d = 1, \ldots, D$; on the other hand, if some $\boldsymbol{\beta}_{r,d} = \mathbf{0}$ we only need to take into account those non-zero vectors in the following procedure. For $r = 1, \ldots, R$, we solve the following optimization problem

$$\begin{aligned} \underset{\rho_{r,1},\ldots,\rho_{r,D}}{\arg\min} \sum_{d=1}^{D} \frac{1}{2}(1-\lambda_2)\|\rho_{r,d}\boldsymbol{\beta}_{r,d}\|_2^2 + \lambda_2\|\rho_{r,d}\boldsymbol{\beta}_{r,d}\|_1 \\ \text{s.t.} \quad \prod_{d=1}^{D} \rho_{r,d} = 1 \quad \text{and} \quad \rho_{r,d} > 0, \end{aligned} \tag{3.6}$$

and use $\hat{\rho}_{r,d}\boldsymbol{\beta}_{r,d}$ to replace $\boldsymbol{\beta}_{r,d}$ in each iterative step of solving (3.5), where $\{\hat{\rho}_{r,d} : r = 1, \ldots, R, d = 1, \ldots, D\}$ is the minimizer of (3.6). This replacement can ensure the objective function decrease (as shown in Proposition 1). In particular, as described in Appendix A.2.2, (3.6) can be transformed to a convex problem. For $\lambda_2 \in (0, 1)$, the Lagrange method and Newton's method can be used to solve (3.6). While for the special boundary cases, i.e., $\lambda_2 \in \{0, 1\}$, we are able to get the closed form solutions

$$\hat{\rho}_{r,d} = \begin{cases} \dfrac{1}{\|\boldsymbol{\beta}_{r,d}\|_1} \displaystyle\prod_{d=1}^{D} \|\boldsymbol{\beta}_{r,d}\|_1^{1/D}, & \text{if } \lambda_2 = 1, \\ \dfrac{1}{\|\boldsymbol{\beta}_{r,d}\|_2} \displaystyle\prod_{d=1}^{D} \|\boldsymbol{\beta}_{r,d}\|_2^{1/D}, & \text{if } \lambda_2 = 0. \end{cases}$$

Proposition 1. *Suppose $\Theta(\boldsymbol{\theta})$ is the scale class of $\boldsymbol{\theta} = (\tilde{\nu}, \mathbf{B}_1, \ldots, \mathbf{B}_D, \tilde{\mathbf{B}}_{D+1})$ up to scaling, i.e.,*

$$\Theta(\boldsymbol{\theta}) = \{\boldsymbol{\theta}^{\boldsymbol{\rho}} : \boldsymbol{\theta}^{\boldsymbol{\rho}} = (\tilde{\nu}, \mathbf{B}_1\boldsymbol{\rho}_1, \ldots, \mathbf{B}_D\boldsymbol{\rho}_D, \tilde{\mathbf{B}}_{D+1}), \boldsymbol{\rho}_d = \mathrm{diag}(\rho_{1,d}, \ldots, \rho_{R,d}), \prod_{d=1}^{D} \rho_{r,d} = 1\}$$

and the solution of (3.6) *is* $\hat{\rho}_{r,d}$, $r = 1, \ldots R$, $d = 1, \ldots, D$. *Let* $\bar{\boldsymbol{\theta}} = (\tilde{\nu}, \bar{\mathbf{B}}_1, \ldots, \bar{\mathbf{B}}_D, \tilde{\mathbf{B}}_{D+1})$, $\bar{\mathbf{B}}_d = (\hat{\rho}_{1,d}\boldsymbol{\beta}_{1,d}, \ldots, \hat{\rho}_{R,d}\boldsymbol{\beta}_{R,d})$, *then*

$$LG(\bar{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}^{\boldsymbol{\rho}} \in \Theta(\boldsymbol{\theta})} LG(\boldsymbol{\theta}^{\boldsymbol{\rho}}).$$

Furthermore, if $\boldsymbol{\beta}_{r,d} \neq \mathbf{0}$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, *then*

$$LG(\bar{\boldsymbol{\theta}}) < LG(\boldsymbol{\theta}^{\boldsymbol{\rho}}), \quad \forall \boldsymbol{\theta}^{\boldsymbol{\rho}} \in \Theta(\boldsymbol{\theta}), \boldsymbol{\theta}^{\boldsymbol{\rho}} \neq \bar{\boldsymbol{\theta}}.$$

Proposition 1 indeed shows that $\bar{\boldsymbol{\theta}}$ is the unique minimizer over $\Theta(\boldsymbol{\theta})$. Although $\Theta(\boldsymbol{\theta})$ is not a convex set, we are able to find the minimizer over $\Theta(\boldsymbol{\theta})$ using the rescaling strategy (3.6).

Algorithm 1: Scale-adjusted block relaxation algorithm.

Input : $\boldsymbol{\theta}^{(0)} = \big(\tilde{\nu}^{(0)}, \mathbf{B}_1^{(0)}, \ldots, \mathbf{B}_D^{(0)}, \mathbf{B}_{D+1}^{(0)}\big)$, $\epsilon > 0$ and $t = 0$.
repeat
  for d *from* $1, \ldots, D, D+1$ **do**
    $\mathbf{B}_d^{(t+1)} = \arg\min_{\mathbf{B}_d} LG(\tilde{\nu}^{(t)}, \mathbf{B}_1^{(t+1)}, \ldots, \mathbf{B}_{d-1}^{(t+1)}, \mathbf{B}_d, \mathbf{B}_{d+1}^{(t)}, \ldots, \mathbf{B}_D^{(t)}, \tilde{\mathbf{B}}_{D+1}^{(t)})$;
  end
  $\tilde{\nu}^{(t+1)} = \arg\min_{\tilde{\nu}} LG\big(\tilde{\nu}, \mathbf{B}_1^{(t+1)}, \ldots, \mathbf{B}_D^{(t+1)}, \tilde{\mathbf{B}}_{D+1}^{(t+1)}\big)$;
  Replace $\mathbf{B}_d^{(t+1)}$ by $\big(\hat{\rho}_{1,d}\boldsymbol{\beta}_{1,d}^{(t+1)}, \ldots, \hat{\rho}_{R,d}\boldsymbol{\beta}_{R,d}^{(t+1)}\big)$, where $\hat{\rho}_{r,d}^{(t+1)}$, $r = 1, \ldots, R$, are obtained from (3.6);
  $t = t + 1$;
until $-LG(\boldsymbol{\theta}^{(t+1)}) + LG(\boldsymbol{\theta}^{(t)}) \leq \epsilon$.
Output: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t)}$.

The above discussion leads us to Algorithm 1 and its convergence property is presented in Proposition 2. This algorithm can be regarded as an improvement version of the Proposition 1 of Zhou et al. (2013), where they required an assumption that the set of stationary points are isolated (modulo permutation and scaling indeterminacy). Even this assumption holds, the scaling indeterminacy can lead to an infinite number of stationary points, which may make the algorithm

unstable in applications. For stability of the algorithm, Zhou et al. (2013) considered a standardization step for both penalized and unpenalized methods. Particularly, their standardization step is in fact a special case of our rescaling strategy for $\lambda_2 = 0$ (l_2 penalty), which can handle the scaling indeterminacy and thus stabilizes the algorithm. Although their standardization step works well in application, for other penalty, such as l_1, this step may increase the value of the objective function and the convergence of the penalized algorithm may not be guaranteed. Using the rescaling strategy (3.6), the convergence property of our proposed penalized algorithm is demonstrated in Proposition 2 and its proof is deferred in Appendix.

Proposition 2. *Assume that the set* $\{\boldsymbol{\theta}, LG(\boldsymbol{\theta}) \leq LG(\boldsymbol{\theta}^{(0)})\}$ *is compact,* $\lambda_1 > 0$*,* $\lambda_2 < 1$ *and the set of stationary points of* $LG(\boldsymbol{\theta})$ *are isolated. Then the sequence* $\boldsymbol{\theta}^{(t)}$ *generated by Algorithm I converges to a stationary point of* $LG(\boldsymbol{\theta})$*.*

3.3 Tunning parameters

The commonly used method to determine the tuning parameters, including the CP rank R, the penalty parameters λ_1 and λ_2, is cross-validation. However, it suffers heavy computation burden in the tensor scenario, especially when the dataset is large. We thus alternatively use the validation method (see, e.g., Chapter 11 of Shalev-Shwartz and Ben-David, 2014) in our numerical experiments, which shows computational attraction.

4. THEORETICAL STUDY

Throughout the theoretical analysis, we assume that the true regression function $m_0(\mathbf{X})$ is a multivariate continuous function and has the following form of representation

$$m_0(\mathbf{X}) = \nu_0 + \frac{1}{s}\sum_{r=1}^{R_0} \langle \boldsymbol{\beta}_{0r,1} \circ \ldots \circ \boldsymbol{\beta}_{0r,D}, F_{0r}(\mathbf{X}) \rangle \,,$$

where $\int_0^1 f_{0,r}(x)\mathrm{d}x = 0$ and $\{f_{0r}\}_{r=1}^{R_0} \subset \mathcal{H}$ is a minimal representation which has been introduced ahead of Theorem 3. $\mathcal{H}$ is the function class that the true broadcasted functions lie in and is specified in Assumption 3. To simplify the notations, we write $\mathbf{B}_{0r} = \boldsymbol{\beta}_{0r,1} \circ \ldots \circ \boldsymbol{\beta}_{0r,D}$ and define a mapping $\mathcal{I} : \mathbb{R}^{p_1 \times \ldots \times p_D \times K} \times \mathbb{R} \to \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$ by

$$\mathbf{A}^\flat = \mathcal{I}(\mathbf{A}, \nu), \tag{4.1}$$

where $\mathbf{A}^\flat_{i_1,\cdots,i_D,k} = \mathbf{A}_{i_1,\cdots,i_D,k}$, for $(i_1, \cdots, i_D) \neq (1, \cdots, 1)$ and $\mathbf{A}^\flat_{1,\ldots,1,k} = \mathbf{A}_{1,\ldots,1,k} + s\nu$, $k = 1, \ldots, K$. It then follows from the property of B-spline functions that

$$\nu + \frac{1}{s}\langle \mathbf{A}, \Phi(\mathbf{X}) \rangle = \frac{1}{s}\langle \mathbf{A}^\flat, \Phi(\mathbf{X}) \rangle. \tag{4.2}$$

As we see in (4.2), the constant ν can be absorbed in the coefficients of B-spline basis of one predictor. This property helps us develop the asymptotic theory according to the fact that the tensor of coefficients inherits the same CP structure. Furthermore, it also goes for other commonly used bases, such as the truncated power basis. Indeed, we will show the asymptotic results of Theorem 1 is also valid for other equivalent bases (see Theorem 4).

4.1 Assumptions

We use C and C with subscripts to refer to generic constants that may change values from context to context. We need the following regularity assumptions.

Assumption 1. *The covariates* $\mathbf{X} \in [0,1]^{p_1 \times \cdots \times p_D}$ *has a continuous density function* g*, which is bounded away from zero and infinity on* $[0,1]^{p_1 \times \cdots \times p_D}$*, i.e., there exist constants* $C_1, C_2 > 0$ *such that* $C_1 \leq g(\mathbf{x}) \leq C_2$ *for all* $\mathbf{x} \in [0,1]^{p_1 \times \cdots \times p_D}$.

Before presenting the assumption related to the random error, we first give the definition of sub-Gaussian random variable and its sub-Gaussian norm.

Definition 1 (sub-Gaussian random variable)**.** *We say that a random variable* X *is sub-Gaussian if the moments satisfies*

$$(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq C\sqrt{p},$$

for any $p \geq 1$ *with a positive constant* C*. The minimum value of* C *is called sub-Gaussian norm of* X*, denoted by* $\|X\|_{\psi_2}$ *(see, for example, Chapter 2.5.2 of Vershynin, 2018).*

Assumption 2. *The vector of random errors,* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\intercal$*, has independent and identically distributed entries. Each* ϵ_i *is mean* 0 *and sub-Gaussian with sub-Gaussian norm* $\sigma < \infty$*.*

Assumption 3. *Let* l *be a nonnegative integer and let* $\tau = l + \omega > 1/2$*, where* $\omega \in (0,1]$*. Let* $\mathcal{H}$ *denote the space of functions on* $[0,1]$ *satisfying the Hölder condition of order* ω*, i.e.,*

$$\begin{aligned} \mathcal{H} = \big\{ g : |g^{(l)}(x_1) - g^{(l)}(x_2)| \leq C|x_1 - x_2|^\omega, \ \forall\, x_1, x_2 \in [0,1], \\ C \in (0,\infty), \int_0^1 g(x)\mathrm{d}x = 0 \big\}, \end{aligned} \tag{4.3}$$

where $g^{(l)}$ *is the* l*-th derivative of* g*. We assume* $f_{0r} \in \mathcal{H}$*,* $r = 1, \ldots, R_0$*.*

Assumption 4. *The order of the B-spline used in* (4.2) *satisfies* $\zeta \geq \tau + \frac{1}{2}$*. We let* $0 = \xi_1 < \xi_2 < \cdots < \xi_{K-\zeta+2} = 1$ *denote the knots of B-spline basis and assume that*

$$h_n = \max_{k=1,\ldots,K-\zeta+1} |\xi_{k+1} - \xi_k| \asymp K^{-1} \quad \textit{and} \quad h_n \Big/ \min_{k=1,\ldots,K-\zeta+1} |\xi_{k+1} - \xi_k| \leq C.$$

Assumptions 1, 3, and 4 are commonly seen in the general nonparametric models. In particular, Assumption 1 ensures the population level of the design matrix has certain eigenvalue property.

In the scope of nonparametric additive models, Stone (1985) and Chen et al. (2018) used this assumption to develop their asymptotic analysis. Assumptions 3 and 4 regularize the space where the true broadcasted functions lie in and guarantee that they can be globally approximated by B-spline functions. Indeed, a well-known result based on these assumptions is that there exist $\boldsymbol{\alpha}_{0,r} = (\alpha_{0r,1}, \ldots, \alpha_{0r,K})^\intercal$, $r = 1, \ldots, R$, such that

$$\left\| f_{0r} - \sum_{k=1}^{K} \alpha_{0r,k} b_k \right\|_\infty = \mathcal{O}(K^{-\tau}), \tag{4.4}$$

where $\|f\|_\infty$ denotes the L_∞ norm of function f. Though we assume $\int_0^1 f_{0r}(x)\mathrm{d}x = 0$, Lemma 6 still implies that there are $\boldsymbol{\alpha}_{0,r}$, $r = 1, \ldots, R$, satisfying (4.4) with

$$\sum_{k=1}^{K} \int_0^1 \alpha_{0r,k} b_k(u)\mathrm{d}u = 0.$$

Despite this mild difference in model identifiability, similar assumptions can be found in Zhou et al. (1998) and Huang et al. (2010). Assumption 2 is recently used in both the regression literature (Wei and Huang, 2010; He and Huang, 2016) and the nonparametric modeling (He et al., 2018). With this assumption, the upper tail probability of the random error is able to be controlled, which slightly generalizes the canonical result of normally distributed noise.

4.2 Convergence rates

We present the convergence rates of $\hat{\mathbf{A}}^\flat$ and $\hat{m}(\mathbf{X})$ in terms of the Hilbert-Schmidt norm and the L_2 norm, respectively. Hilbert-Schmidt norm is a generalization from Frobenius norm of matrices to tensors, which is defined as $\|\mathbf{A}\|_{HS} = \langle \mathbf{A}, \mathbf{A} \rangle^{1/2}$ for any generic tensor $\mathbf{A}$. While L_2 norm is defined as $\|f(\mathbf{X})\|_{L_2} = [\{\mathbb{E}_{\mathbf{X}} f(\mathbf{X})\}]^{1/2}$ for any function $f \in \mathcal{H}$. We also denote $\mathbf{A}_0$ as

$$\mathbf{A}_0 = \sum_{r=1}^{R_0} \mathbf{B}_{0r} \circ \boldsymbol{\alpha}_{0r},$$

where $\boldsymbol{\alpha}_{0r}$ satisfies (4.4) under mean zero constraint, $r = 1, \ldots, R$. Theorem 1 shows the convergence rates of the unpenalized estimators, where the parameters p_i, K, R and R_0 are allowed to go to infinity with the sample size n.

Theorem 1. *Suppose* $(\hat{\mathbf{A}}, \hat{\nu})$ *is a solution of* (3.3) *and* $\hat{m}(\mathbf{X})$ *is the corresponding estimated regression function. Let* $\hat{\mathbf{A}}^\flat = \mathcal{I}(\hat{\mathbf{A}}, \hat{\nu})$ *and* $\mathbf{A}_0^\flat = \mathcal{I}(\mathbf{A}_0, \nu_0)$*. If Assumptions 1–4 hold,* $R \geq R_0$*, and* $n > C\tilde{h}_n^2 h_n^{-2}\big(R^{D+1} + \sum_{i=1}^D Rp_i + RK\big)$ *for some* $C > 0$*, then we have the following results*

i.

$$\begin{aligned}\frac{1}{\sqrt{s}}\|\hat{\mathbf{A}}^\flat - \mathbf{A}_0^\flat\|_{HS} \\ = \mathcal{O}_{\mathrm{p}}\left(\left[\frac{sK\big\{R^{D+1} + \sum_{i=1}^D Rp_i + RK\big\}}{n}\right]^{1/2}\right) \\ + \mathcal{O}_{\mathrm{p}}\left(\left\{\frac{\sum_{r=1}^{R_0}\|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{\sqrt{s}}\right\}\frac{1}{K^{\tau-1/2}}\right);\end{aligned} \tag{4.5}$$

ii.

$$\begin{aligned}\|\hat{m}(\mathbf{X}) - m_0(\mathbf{X})\|_{L_2}^2 \\ = \mathcal{O}_{\mathrm{p}}\left(\frac{R^{D+1} + \sum_{i=1}^D Rp_i + RK}{n}\right) \\ + \mathcal{O}_{\mathrm{p}}\left(\left\{\frac{\sum_{r=1}^{R_0}\|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s}\right\}^2\frac{1}{K^{2\tau}}\right),\end{aligned} \tag{4.6}$$

where

$$\tilde{h}_n = \max\left\{\frac{h_n^{1/(-\log h_n)}}{(-2\log h_n)}, h_n\right\}.$$

Roughly speaking, the first term and the second term in (4.6) correspond to the estimation error and the approximation error, respectively. The condition $R \geq R_0$ is used to bound the approximation error from above. Without this condition, the estimated function will converge to the best R rank approximation. The condition on n ensures the difference between the eigenvalues of the gram matrix of "design" in population level and its empirical counterpart is negligible, compared with the rates of convergence.

Remark 1. The proof of Theorem 1 is not straightforward even if we discard the low-rank and

broadcasting structure of the proposed model (2.2). To see this, we can vectorize the basis tensor and its coefficients in (3.2), and reconstruct the regression function as the nonparametric additive model. The main challenge of studying the convergence rates is to determine the upper and lower bounds for the eigenvalues of the gram matrix of "design". Many existing works, such as Ravikumar et al. (2009), directly assume the eigenvalues are bounded away from zero and infinity. It is, however, not clear to be true in general, since the number of basis functions goes to infinity with the sample size in order to guarantee the proper approximation property. Indeed, it can be proved that such assumption fails for B-spline basis when there is a divergent number of additive component functions. When the number of additive component functions is a fixed constant, Huang et al. (2010) shows the bounds of the eigenvalues using Lemma 3 of Stone (1985) and Lemma 6.2 of Zhou et al. (1998). It is worth mentioning that directly using the results of Stone (1985) will lead the convergence result at an exponential rate when the number of additive component functions goes to infinity with the sample size n (see, e.g., Chen et al., 2018). Therefore, Theorem 1 fills in the gap to allow the number of additive component functions to diverge. Furthermore, we incorporate the local Gaussian width arguments of Banerjee et al. (2015) and the covering number arguments of Rauhut et al. (2017) to overcome the difficulties due to the low-rank and broadcasting structure of (2.2).

For different combinations of orders between the parameters (R, R_0, p_i) and the sample size n, we can tune the number of basis functions K to get the optimal rates of convergence. Let

$$\delta_1 = R^{D+1} + \sum_{i=1}^{D} R p_i \quad \text{and} \quad \delta_2 = \left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\}^2.$$

If $\delta_1 \delta_2^{-1/(2\tau+1)} R^{-2\tau/(2\tau+1)} \geq n^{1/(2\tau+1)}$, the optimal rate can be tuned is δ_1/n when K satisfies $(n\delta_2/\delta_1)^{1/2\tau} \lesssim K \lesssim \delta_1/R$. On the other hand, if $\delta_1 \delta_2^{-1/(2\tau+1)} R^{-2\tau/(2\tau+1)} < n^{1/(2\tau+1)}$, letting $K \sim (n\delta_2/R)^{1/(2\tau+1)}$ will lead the optimal rate of convergence to $(R/n)^{2\tau/(2\tau+1)} \delta_2^{1/(2\tau+1)}$. One special case is that when p_i, R and R_0 do not grow with n, choosing $K \sim n^{1/(2\tau+1)}$ is able to obtain the optimal rate of convergence $n^{-2\tau/(2\tau+1)}$ as in Stone (1982). Theorem 1 indeed generalizes the

canonical results to tensor low-rank modeling with broadcasting.

Although Theorem 1 guarantees the asymptotic performance of the unpenalized estimators, in many real applications the penalized estimation is needed, especially when the number of predictors are moderately large. Theorem 2 shows the rates of convergence of the penalized method in terms of concentration inequalities. Suppose

$$\mathbf{B}_{0r} = \boldsymbol{\beta}_{0r,1} \circ \cdots \circ \boldsymbol{\beta}_{0r,D}, \quad r = 1, \ldots, R_0,$$

are the rank-one decomposition that make the penalty term in (3.5) smallest over all such decompositions, which is well-defined according to Proposition 1. For simplicity, we denote

$$G_0 = \sum_{r=1}^{R_0} \sum_{d=1}^{D} \sum_{i=1}^{p_d} P_\lambda(\beta_{0r,di_d}). \tag{4.7}$$

Similar to Theorem 1, p_i, K, R and R_0 are allowed to go to infinity with the sample size n in Theorem 2.

Theorem 2. *Suppose* $(\hat{\mathbf{A}}_p, \hat{\nu}_p)$ *is a solution to* (3.5) *and* $\hat{m}_p(\mathbf{X})$ *is the corresponding estimated regression function. Let* $\hat{\mathbf{A}}_p^\flat = \mathcal{I}(\hat{\mathbf{A}}_p, \hat{\nu}_p)$ *and* $\mathbf{A}_0^\flat = \mathcal{I}(\mathbf{A}_0, \nu_0)$*. If Assumptions 1–4 hold,* $R \geq R_0$ *and* $n > C\tilde{h}_n^2 h_n^{-2}\big(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK\big)$ *for some* $C > 0$*, then*

i.

$$\frac{1}{\sqrt{s}}\|\hat{\mathbf{A}}_p^\flat - \mathbf{A}_0^\flat\|_{HS} \leq \frac{\{s\delta_3^2 + (4sKG_0)/n\}^{1/2} + \sqrt{s}\delta_3}{2}; \tag{4.8}$$

ii.

$$\|\hat{m}_p(\mathbf{X}) - m_0(\mathbf{X})\|_{L_2}^2 \leq \frac{C_1\{\delta_3^2 + (4KG_0)/n\}}{K} \tag{4.9}$$

with probability at least

$$1 - C_2 \exp\left\{ - C_3\left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK\right)\right\},$$

where G_0 is defined in (4.7) *and*

$$\delta_3 = C_4\left\{\frac{K\left(R^{D+1}+\sum_{i=1}^{D}Rp_i+RK\right)}{n}\right\}^{1/2}+C_5\left\{\frac{\sum_{r=1}^{R_0}\|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s}\right\}\frac{1}{K^{\tau-1/2}}.$$

Compared with Theorem 1, Theorem 2 has an addition term G_0, which is the bias due to the elastic net penalty. When the penalty function is small relatively to the estimation and approximation errors, this bias can be negligible in the viewpoint of rates of convergence. On the other hand, though the penalized estimators may have slower rates of convergence than the unpenalized ones, it will stabilize the performance of estimation and lead to a parsimonious model such that some regions will be identified as irrelevance.

5. EXPERIMENTS AND CONCLUSIONS

To confirm the effectiveness of the broadcast nonparametric tensor regression (BNTR) method, we investigate the performance on both synthetic and real data, and compare with (i) Tensor Linear Regression (TLR, Zhou et al., 2013), and (ii) Elastic Net Regression on the vectorized tensor predictor (ENetR, Zou and Hastie, 2005). For ENetR, we use the R package glmnet (Hastie and Qian, 2014). For TLR, we use the benchmark MATLAB TensorReg toolbox (Zhou et al., 2013). Since our rescaling strategy can also enhance the algorithm of TLR implemented in the TensorReg toolbox, for a relatively pair comparison, we also consider the TLR algorithm with the rescaling strategy. To distinguish the two algorithm for TLR, we use TLR-1 and TLR-2 to represent the algorithm of Zhou et al. (2013) and our improvement algorithm, respectively. For BNTR, similar to Huang et al. (2010), we use the cubic spline and fix the number of basis $K = 7$. The knots are chosen as the equally spaced quantiles.

There are two aims, i.e., confirming the advantages of BNTR in region selection and prediction tasks. Unlike TLR, the important region for BNTR can not be identified directly using the estimated coefficient tensor $\hat{\mathbf{A}}_p$, since the contribution for each pixel of the input contains in the mode-$(D+1)$ fiber (the higher-order analogue of matrix rows and columns) (Kolda and Bader, 2009) of $\hat{\mathbf{A}}_p$. To summarize the contribution from each pixel, we consider a norm tensor

$$\mathbf{B}_f \in \mathbb{R}^{p_1 \times \cdots \times p_D}, \tag{5.1}$$

where $(\mathbf{B}_f)_{i_1,\ldots,i_D} = \|\hat{f}_{i_1,\ldots,i_D}\|_2 = \{\int_0^1 \hat{f}^2_{i_1,\ldots,i_D}(x)\mathrm{d}x\}^{1/2}$, $\hat{f}_{i_1,\ldots,i_D}(x) = \sum_{k=2}^K \hat{A}_{p,i_1,\ldots,i_D,k} b'_k(x) - \sum_{k=2}^K \hat{A}_{p,i_1,\ldots,i_D,k} \int b'_k(x)\mathrm{d}x$, $\hat{A}_{p,i_1,\ldots,i_D,k}$ is the $(i_1,\ldots,i_D,k)$-th entry of $\hat{\mathbf{A}}_p$, $i_1 = 1,\ldots,p_1,\ldots$, $i_D = 1,\ldots,p_D$, and $\{b'_k(x)\}_{k=2}^K$ is the truncated power basis without the constant one. We use this norm tensor (5.1) to identify important sub-regions. This is a simple paradigm that works in application. In specific applications, subject preference may prefer alternative paradigms. To measure the performance for prediction, since we know the true function in the synthetic data but

do not know that of the real data, we use the mean integrated squared error (MISE) in the synthetic data while the mean squared prediction error (MSPE) in the real data. The MISE is defined as

$$\text{MISE} = \|\hat{m}(\mathbf{X}) - m(\mathbf{X})\|_{L_2}^2,$$

which can be numerically calculated after decomposition to the sum of entry functions, while the MSPE is defined as

$$\text{MSPE} = \frac{1}{n_t}\sum_{i=1}^{n_t}(y_i - \hat{y}_i)^2, \tag{5.2}$$

where n_t is the test sample size, $\hat{y}_i$ is the prediction value and y_i is the observed value in the test set. Note that MISE is a more precise quantity to measure the performance, and MSPE is commonly used in real world. We will report selected results from synthetic examples and applications to the publicly facial data and monkey brain data.

5.1 Synthetic data

For the purpose of illustration, similar to Zhou et al. (2013), we consider $\mathbf{X} \in \mathbb{R}^{64\times 64}$ in this section. These simulation results demonstrate that BNTR and TLR-1,2 are comparable in the low-rank linear setting, whereas BNTR is much better than the linear models (TLR-1,2, and ENetR) when the data involves nonlinear relationship.

5.1.1 Data generation

We consider 4 different data generation procedures, i.e.,

$$\text{Case 1.} \quad y = m_1(\mathbf{X}) + \epsilon_1 = 1 + \langle \mathbf{B}_1, \mathbf{X}\rangle + \epsilon_1,$$

$$\text{Case 2.} \quad y = m_2(\mathbf{X}) + \epsilon_2 = 1 + \langle \mathbf{B}_2, F_1(\mathbf{X})\rangle + \epsilon_2,$$

$$\text{Case 3.} \quad y = m_3(\mathbf{X}) + \epsilon_3 = 1 + \langle \mathbf{B}_{31}, F_1(\mathbf{X})\rangle + \langle \mathbf{B}_{32}, F_1(\mathbf{X})\rangle + \epsilon_3,$$

$$\text{Case 4.} \quad y = m_4(\mathbf{X}) + \epsilon_4 = 1 + \langle \mathbf{B}_{41}, F_1(\mathbf{X})\rangle + \langle \mathbf{B}_{42}, F_2(\mathbf{X})\rangle + \epsilon_4,$$

where the broadcasted function F_1 and F_2 satisfying

$$(F_1(\mathbf{X}))_{i_1,i_2} = f_1(X_{i_1,i_2}) = X_{i_1,i_2} + 0.6\sin(2\pi(X_{i_1,i_2} - 0.5)^2),$$

and

$$(F_2(\mathbf{X}))_{i_1,i_2} = f_2(X_{i_1,i_2}) = X_{i_1,i_2} + 0.3\cos(2\pi X_{i_1,i_2}),$$

for $i_1 = 1, \ldots, 64$, $i_2 = 1, \ldots, 64$. The true signal $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_{31}$, $\mathbf{B}_{32}$, $\mathbf{B}_{41}$ and $\mathbf{B}_{42}$ are binary with the true signal sub-region equals to one and the rest zero, the input $\mathbf{X}$ has standard uniform distribution entries, $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2, 3, 4$ and σ_j is used for adjusting the signal level. These cases demonstrate four different situations, i.e,

(1) low rank linear model with one important sub-region,

(2) low rank nonlinear model with one important sub-region,

(3) low rank nonlinear model with two separated important sub-regions that share the same nonlinearity,

(4) low rank nonlinear model with two separated important sub-regions that share different nonlinearities.

For each cases, we randomly generate a set of independent samples under the signal strength $\sigma_j = 10\%$ of the standard deviation of $m_j(\mathbf{X})$, $j = 1, 2, 3, 4$, and split the sample set into two separate subsets, i.e., the validation set with 20% data and the training set with 80% data. We train the models in the training set and tune the tunning parameters in the validation set. For the grid of tuning parameters in each simulated experiment, we consider all combinations of $R \in \{1, 2, 3, 4, 5\}$, $\lambda_1 \in \{10^{-2}, 5 \times 10^{-1}, 10^{-1}, \ldots, 10^2, 5 \times 10^2, 10^3\}$ and $\lambda_2 \in \{0, 0.5, 1\}$. We summarize our findings in later two sub-sections.

5.1.2 Identifying important sub-regions

The important sub-regions for BNTR are identified from the norm matrix $\mathbf{B}_f \in \mathbb{R}^{64\times 64}$, which is defined in (5.2), while that of TLR-1 and TLR-2 are from the estimated coefficients. We rescale the norm matrix and estimated coefficients to $[0, 1]$ and use $rasterImage$ function in R to implement them. The results for the comparison among TLR-1, TLR-2 and BNTR in the sample size $n = 1000$, are shown in Figure 5.1, from which we can see that BNTR and TLR-1,2 have similar region selection result in terms of Case 1 (the low rank model without nonlinearity), whereas BNTR is much better than TLR-1,2 for Case 2, 3 and 4 (the low rank model with nonlinearity).

Although we compare the models under the sample size $n = 1000$, it is not the minimal sample size that is needed to identify the important sub-regions. To demonstrate this, we also report the results of BNTR for various sample sizes ($n = 500, 750$, or 1000) in Figure 5.2. The minimal sample size needed to identify the important region varies in different true signals. The signal with lower degree of complexity, e.g., Case 1, can be found the important sub-region with a small sample size, while these signals with higher degree of complexity, e.g., Case 2, 3 and 4, need more samples.

Remark 2. Note that the black point in these region selection figures (including Figure 5.1, 5.2 and 5.3) may not zero exactly, but some small numbers. If one wants more sparse solutions, the thresholding can be applied.

5.1.3 Estimation performance

We consider the estimation performance with varying sample size $n \in \{500, 750, 1000\}$, where 20% data are used for validation. The results are evaluated based on 50 replications and shown in Table 5.1. Overview, the results of BNTR confirm our theory. Particularly, for the nonlinear situations (Case 2, 3, and 4), it can be found that BNTR is much better than all the linear models, which demonstrates the excellent performance of BNTR. For the linear Case 1, we would see that BNTR is also very good, which demonstrates the advantages of economically modeling idea mentioned in Section 2. Besides, we found that TLR-2 is better than TLR-1, which shows that the

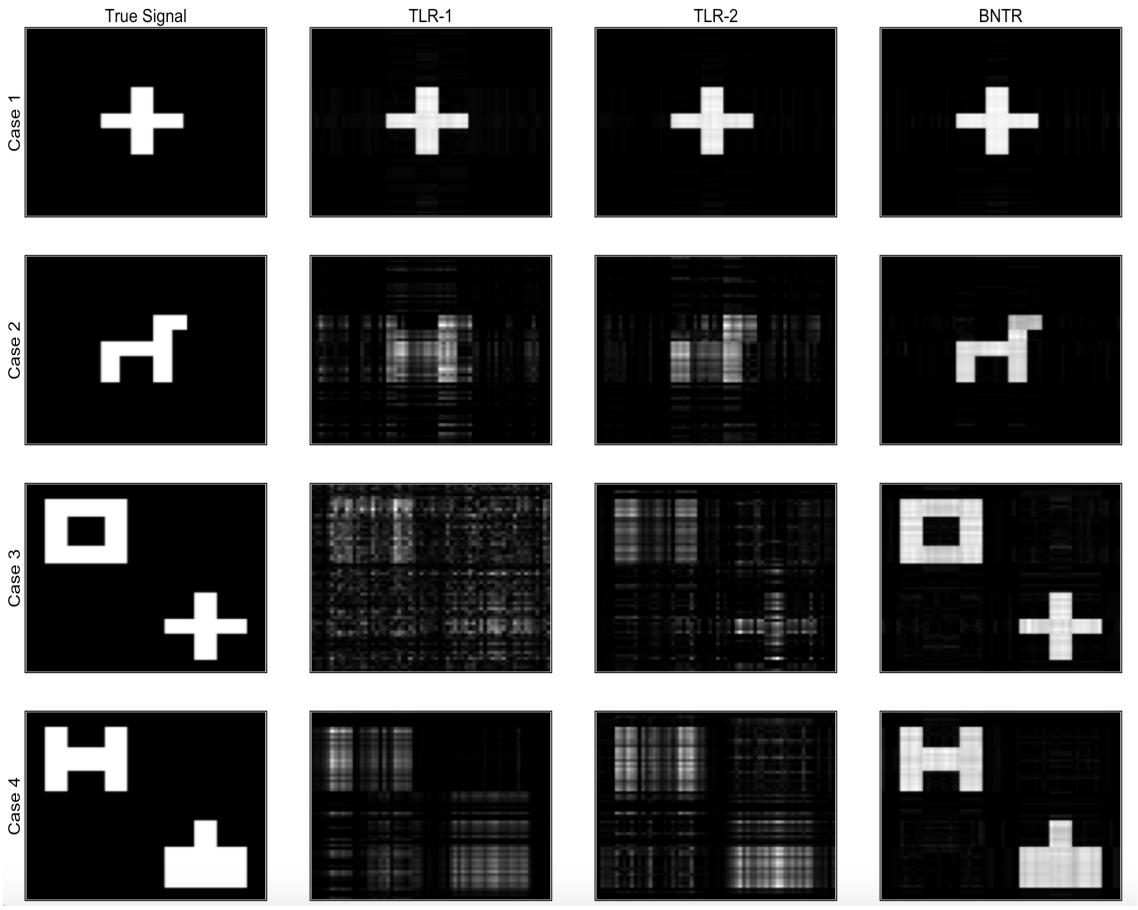

Figure 5.1: Region selection comparison among TLR-1, TLR-2 and BNTR when the total sample size $n = 1000$ and 20% data are used for validation. Here the true signals are $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_{31} + \mathbf{B}_{32}$ and $\mathbf{B}_{41} + \mathbf{B}_{42}$ for Case 1, 2, 3 and 4, respectively.

rescaling strategy is not only a theoretical guarantee for the convergence of the algorithm, but also an improvement in practice.

5.2 Real data

We also examine the performance of our method on two publicly available benchmark data sets for tensor regression application. In the real data analysis, we consider more various values for the rank R and penalized parameters λ_1, i.e., $R \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $\lambda_1 \in \{10^{-2}, 2.5 \times 10^{-2}, 5 \times 10^{-2}, 7.5 \times 10^{-2}, 10^{-1}, \ldots, 10^{2}, 2.5 \times 10^{2}, 5 \times 10^{2}, 7.5 \times 10^{2}, 10^{3}\}$.

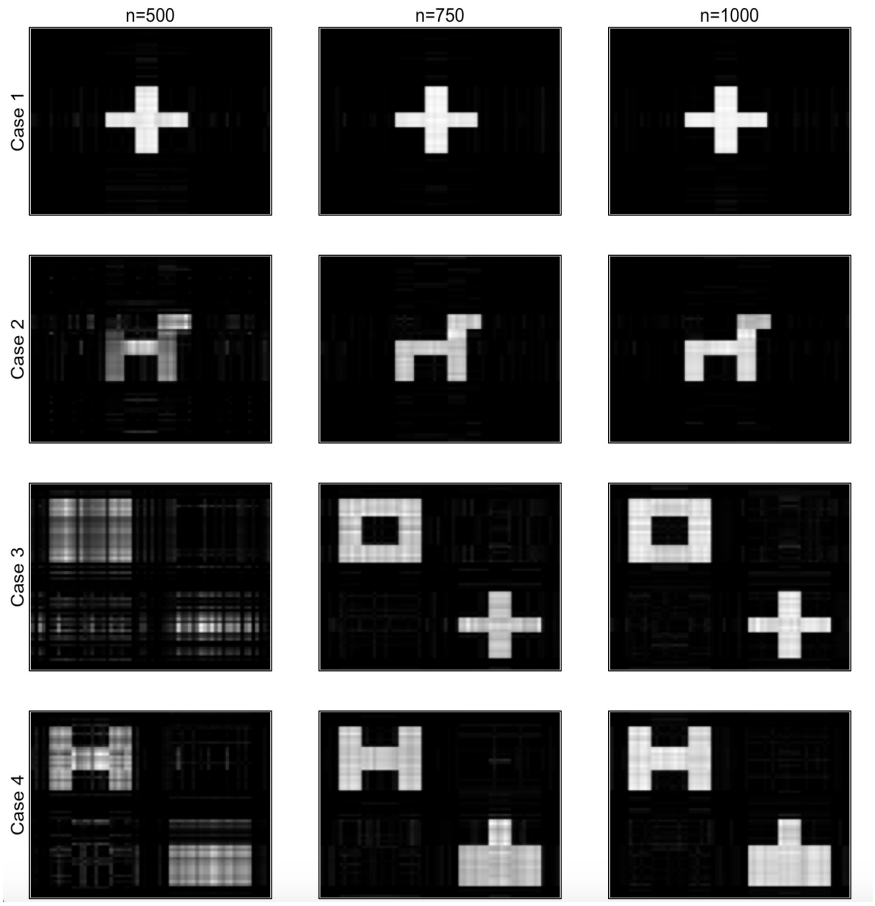

Figure 5.2: Region selection of BNTR for Case 1, 2, 3 and 4, in various sample size ($n = 500, 750, 1000$), where 20% data are used for validation.

5.2.1 Facial data

We apply our model to facial images of the Labeled Faces in the Wild database (Learned-Miller et al., 2016), which is also analyzed by a tensor-on-tensor regression method (Lock, 2018). There are about 13000 publicly available images taken from the internet, and 73 describable attributes (Kumar et al., 2009) for each facial image. The attributes are measured continuously, and the higher values, the more obvious attribute. The goal is to predict the attribute based on the facial images, which helps to describe images and further study (e.g., see Farhadi et al., 2009). For the output, we consider one example, i.e, attribute 22, harsh lighting. Intuitively, this attribute may imply the situation information. For the input, we follow the data preprocessing procedure described in

Table 5.1: Estimation comparison in the synthetic data. Reported are mean MISE and its standard deviation (in parenthesis) based on 50 data replications. Here n is the total sample size and 20% of the data will be used for validation.

n	Case	TLR-1	TLR-2	ENetR	BNTR
500	1	0.2227 (0.0612)	$\mathbf{0.0655}(0.0103)$	16.55 (0.5619)	0.0902 (0.0182)
	2	24.51 (3.323)	22.18 (2.024)	31.33 (0.7324)	$\mathbf{3.182}(1.337)$
	3	77.18 (8.909)	75.87 (6.967)	75.87 (2.153)	$\mathbf{26.06}(5.766)$
	4	92.51 (10.12)	92.86 (5.576)	89.69 (2.629)	$\mathbf{23.29}(5.478)$
750	1	0.1077 (0.0228)	$\mathbf{0.0403}(0.0056)$	14.75 (0.4567)	0.0548 (0.0076)
	2	20.42 (2.020)	17.50 (1.302)	30.52 (0.5072)	$\mathbf{0.7616}(0.2773)$
	3	75.28 (12.01)	56.17 (4.975)	74.12 (2.396)	$\mathbf{3.965}(3.240)$
	4	86.47 (10.28)	64.35 (4.432)	87.26 (2.574)	$\mathbf{4.402}(2.276)$
1000	1	0.0781 (0.0139)	$\mathbf{0.0291}(0.0031)$	10.61 (0.6172)	0.0395 (0.0052)
	2	17.84 (1.831)	14.99 (0.5938)	29.75 (0.4986)	$\mathbf{0.3323}(0.0589)$
	3	69.86 (12.31)	45.82 (2.251)	72.35 (2.412)	$\mathbf{0.9691}(0.1685)$
	4	66.30 (7.771)	54.68 (2.127)	84.49 (2.173)	$\mathbf{1.4106}(0.4695)$

(Lock, 2018) and get a $90 \times 90 \times 3$ input tensor for each image, where the components of face are located in similar positions for different images. We randomly choose 2000 different images of the whole data set and randomly split them to 3 different set, i.e., 1000 images in the training set, 500 images in the validation set and 500 images in the test set. Note that the norm tensor of this data set is a mode-3 tensor, and the third dimension corresponds to colors. We can identify the important sub-regions roughly in terms of the facial position by transferring the $90 \times 90 \times 3$ norm tensor (5.1) to a matrix of size 90×90, where each entry corresponds to a position. We take l_2 norm of the fiber along the color dimension to achieve this transformation. We report the important sub-region results in Figure 5.3 and the prediction performance comparison among different models in Table 5.2. From Figure 5.3, we could see the sub-region around the eyes is related to harsh lighting, which is consistent with our intuitive understanding. Usually, we can recognize the harsh lighting by the squinty eyes. Thus, one important sub-region should be around the eyes. When squinty eyes, the sub-region around the nose may have some shape change, which may cause the reason why this sub-region is also important. From the prediction performance in Table 5.2, BNTR is much better than TLR-1,2 and ENetR, which implies the attribute 22 have relatively strong nonlinearity

with the input face. Besides, the prediction for TLR-2 is better than TLR-1, which is similar to the situation in synthetic data. This fact also shows that our rescaling strategy is helpful for this entry-wise penalized regression algorithm.

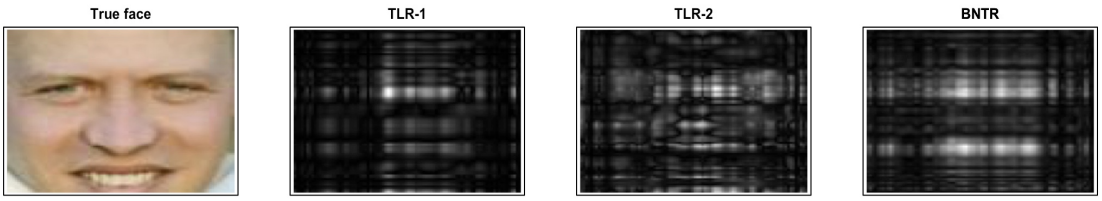

Figure 5.3: Important sub-regions comparison in the facial data.

Table 5.2: Prediction comparison in the facial data. Reported are mean MPSE and its standard deviation (in parenthesis) based on 10 data replications.

Data	TLR-1	TLR-2	ENetR	BNTR
Facial	0.5960 (0.0430)	0.5857 (0.0445)	0.5805 (0.0429)	**0.3207** (0.0332)

5.2.2 Monkey data

We also apply our model to a publicly available benchmark data set for tensor regression application, i.e., the monkey's electrocorticography (ECoG) data (http://neurotycho.org/food-tracking-task). This data is also analyzed by a nonlinear tensor regression model (Hou et al., 2015). Here the input is the preprocessed ECoG signal, which is organized as a three order tensor (channel $\times$ frequency $\times$ time) and the output is the movement distance of the monkey's limb on different 3 markers along each axis (x, y or z). For the data preprocessing of input, the channels are down-sampled to 5 channels in Hou et al. (2015), since their model will suffer the curse of dimensionality and can not handle a higher dimensional input. We do not need to do this down-sample step due to the economical broadcasted nonlinear setting. Our data preprocessing procedure is

similar to Chao et al. (2010) and Shimoda et al. (2012). First, the signals were band-pass filtered from 0.3 to 499Hz and re-referenced using a common average reference montage; then, the time-frequency representation of brain signals at each electrode was described by a scalogram generated by Morlet wavelet transformation at ten different center frequencies (20Hz, 30Hz, ..., 110 Hz); the scalogram of time t was calculated from $t - 1\,\mathrm{s}$ to t and then resampled at 10 time lags, i.e., $t - 900\,\mathrm{ms}, t - 800\,\mathrm{ms}, \ldots, t - 100\,\mathrm{ms}, t$. After a standardization step (z-score) at each frequency over the 10 time lags for each electrode, we get our input tensor of size $64 \times 10 \times 10$. We follow Hou et al. (2015) and choose a subsegment of the whole 15 minutes dataset starting from the 2nd minute comprising 10000 data pairs where the motion data (output) corresponding to the left shoulder marker along the x-axis. And we randomly split these data pairs to 3 different sets, i.s., a training set with size of 4000, a validation set with the size of 1000, and a test set of size 5000. Compare with the aforementioned application in facial data, the training size for the monkey data is bigger, which helps to overcome the estimation error and reveal the approximation error. Since the important sub-regions may vary with time, we do not go further to find important sub-regions in this kind of data. Alternatively, we focus on the prediction performance. We repeat the experiment 10 times and report the prediction results in Table 5.3. We can see that the result of BNTR is the best, which reveals that there is a strong nonlinear relationship between the ECoG data and movements of the monkey.

Table 5.3: Prediction comparison in the monkey data. Reported are mean MPSE and its standard deviation (in parenthesis) based on 10 data replications.

Data	TLR-1	TLR-2	ENetR	BNTR
Monkey	3.1703 (0.0418)	3.0886 (0.0621)	3.1256 (0.0431)	$\mathbf{2.5687}(0.0756)$

REFERENCES

Alizadeh, F. and Goldfarb, D. (2003) Second-order cone programming. *Mathematical Programming*, **95**, 3–51.

Alquier, P. and Biau, G. (2013) Sparse single-index model. *Journal of Machine Learning Research*, **14**, 243–280.

Banerjee, A., Chen, S., Fazayeli, F. and Sivakumar, V. (2015) Estimation with norm regularization. *arXiv preprint arXiv:1505.02294*.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. et al. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, **3**, 1–122.

Boyd, S. and Vandenberghe, L. (2004) *Convex optimization*. Cambridge, UK: Cambridge University Press.

Chandrasekaran, V., Recht, B., Parrilo, P. A. and Willsky, A. S. (2012) The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, **12**, 805–849.

Chao, Z. C., Nagasaka, Y. and Fujii, N. (2010) Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey. *Frontiers in Neuroengineering*, **3**, 3.

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

Chen, Z., Fan, J. and Li, R. (2018) Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, **113**, 315–327.

De Boor, C. (1973) The quasi-interpolant as a tool in elementary polynomial spline theory. *Approximation Theory*, 269–276.

— (1976) Splines as linear combinations of b-splines. a survey. *Tech. rep.*, Wisconsin Univ Madison Mathematics Research Center.

De Lathauwer, L. (2006) A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications*, **28**,

642–666.

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

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

Fan, J., Feng, Y. and Song, R. (2011) Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, **106**, 544–557.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**, 1348–1360.

Farhadi, A., Endres, I., Hoiem, D. and Forsyth, D. (2009) Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1778–1785. IEEE.

Harshman, R. A. (1984) Data preprocessing and the extended parafac model. *Research Methods for Multi-mode Data Analysis*, 216–284.

Hastie, T. and Qian, J. (2014) Glmnet vignette. *Retrieve from http://www. web. stanford. edu/˜ hastie/Papers/Glmnet_Vignette. pdf. Accessed September*, **20**, 2016.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: CRC Press.

He, K. and Huang, J. Z. (2016) Asymptotic properties of adaptive group lasso for sparse reduced rank regression. *Stat*, **5**, 251–261.

He, K., Lian, H., Ma, S. and Huang, J. Z. (2018) Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. *Journal of the American Statistical Association*, **113**, 746–754.

Hoff, P. D. (2015) Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, **9**, 1169.

Horowitz, J. L. and Härdle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, **91**, 1632–1640.

Hou, M., Wang, Y. and Chaib-draa, B. (2015) Online local gaussian process for tensor-variate

regression: Application to fast reconstruction of limb movements from brain signal. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5490–5494. IEEE.

Hu, W., Kong, D. and Shen, W. (2019) Nonparametric matrix response regression with application to brain imaging data analysis. *arXiv preprint arXiv:1904.00495*.

Huang, J., Horowitz, J. L. and Wei, F. (2010) Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**, 2282.

Ichimura, H. (1993) Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, **58**, 71–120.

Imaizumi, M. and Hayashi, K. (2016) Doubly decomposing nonparametric tensor regression. In *International Conference on Machine Learning*, 727–736.

Kanagawa, H., Suzuki, T., Kobayashi, H., Shimizu, N. and Tagami, Y. (2016) Gaussian process nonparametric tensor estimator and its minimax optimality. In *International Conference on Machine Learning*, 1632–1641.

Kang, J., Reich, B. J. and Staicu, A.-M. (2018) Scalar-on-image regression via the soft-thresholded gaussian process. *Biometrika*, **105**, 165–184.

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

Koltchinskii, V. (2011) *Oracle inequalities in empirical risk minimization and sparse recovery problems*. New York: Springer Science & Business Media.

Kruskal, J. B. (1977) Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, **18**, 95–138.

— (1989) Rank, decomposition, and uniqueness for 3-way and n-way arrays. *Multiway Data Analysis*, 7–18.

Kumar, N., Berg, A. C., Belhumeur, P. N. and Nayar, S. K. (2009) Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, 365–372. IEEE.

Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H. and Hua, G. (2016) Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, 189–248. Springer.

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

Lin, Y., Zhang, H. H. et al. (2006) Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34**, 2272–2297.

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

Lu, H., Plataniotis, K. N. and Venetsanopoulos, A. (2013) *Multilinear subspace learning: dimensionality reduction of multidimensional data*. Boca Raton, Florida: Chapman and Hall/CRC.

Meier, L., Van De Geer, S. and Bühlmann, P. (2009) High-dimensional additive modeling. *The Annals of Statistics*, **37**, 3779–3821.

Miranda, M. F., Zhu, H. and Ibrahim, J. G. (2018) Tprm: Tensor partition regression models with applications in imaging biomarker detection. *The Annals of Applied Statistics*, **12**, 1422–1450.

Radchenko, P. (2015) High dimensional single index models. *Journal of Multivariate Analysis*, **139**, 266–282.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012) Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, **13**, 389–427.

Raskutti, G., Yuan, M., Chen, H. et al. (2019) Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, **47**, 1554–1584.

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

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 1009–1030.

Reiss, P. T. and Ogden, R. T. (2010) Functional generalized linear models with images as predic-

tors. *Biometrics*, **66**, 61–69.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric regression*. Cambridge: Cambridge University Press.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding machine learning: From theory to algorithms*. New York: Cambridge University Press.

Shimoda, K., Nagasaka, Y., Chao, Z. C. and Fujii, N. (2012) Decoding continuous three-dimensional hand trajectories from epidural electrocorticographic signals in japanese macaques. *Journal of Neural Engineering*, **9**, 036015.

Sidiropoulos, N. D. and Bro, R. (2000) On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **14**, 229–239.

Stegeman, A. and Sidiropoulos, N. D. (2007) On kruskal's uniqueness condition for the cande-comp/parafac decomposition. *Linear Algebra and its Applications*, **420**, 540–552.

Stone, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 1040–1053.

— (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, **3**, 689–705.

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

Talagrand, M. (2005) *The generic chaining*. Berlin: Springer.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.

Vershynin, R. (2018) *High-dimensional probability: An introduction with applications in data science*. New York: Cambridge University Press.

Wang, F., Zhang, P., Qian, B., Wang, X. and Davidson, I. (2014) Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 145–154. ACM.

Wang, X., Zhu, H. and Initiative, A. D. N. (2017) Generalized scalar-on-image regression models

via total variation. *Journal of the American Statistical Association*, **112**, 1156–1168.

Wei, F. and Huang, J. (2010) Consistent group selection in high-dimensional linear regression. *Bernoulli*, **16**, 1369–1384.

Wood, S. (2006) *Generalized additive models: an introduction with R*. Boca Raton: CRC press.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zhao, Q., Zhang, L. and Cichocki, A. (2013) A tensor-variate gaussian process for classification of multidimensional structured data. In *Twenty-seventh AAAI Conference on Artificial Intelligence*.

Zhao, Q., Zhou, G., Zhang, L. and Cichocki, A. (2014) Tensor-variate gaussian processes regression and its application to video surveillance. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1265–1269. IEEE.

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

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

Zhou, S., Shen, X., Wolfe, D. et al. (1998) Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26**, 1760–1782.

Zhu, Z., Hu, X. and Caverlee, J. (2018) Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1153–1162. ACM.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

APPENDIX A

TECHNICAL RESULTS

To simplify the notations, we let

$$\mathcal{J} = \{(i_1, \cdots, i_D), i_1 = 1, \ldots, p_1, \ldots, \ldots, i_D = 1, \ldots, p_D\}.$$

Note that $s = \Pi_{d=1}^{D} p_d$, then the cardinality $|\mathcal{J}| = s$.

The concept of Gaussian width (Chandrasekaran et al., 2012; Vershynin, 2018) and γ-functionals (Talagrand, 2005; Banerjee et al., 2015) will be used in several places of our proofs. We put their definitions in the beginning of technical results.

Definition 2 (Gaussian width)**.** *For any set $\mathcal{P} \subset \mathbb{R}^p$, the Gaussian width of the set $\mathcal{P}$ is defined as*

$$w(\mathcal{P}) = \mathbb{E}_{\mathbf{x}} \sup_{\mathbf{a} \in \mathcal{P}} \langle \mathbf{a}, \mathbf{x} \rangle,$$

where the expectation is over $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_{p \times p})$, a vector of independently standard Gaussian random variables.

Definition 3 (γ-functionals)**.** *Consider a metric space (T, d) and for a finite set $\mathcal{A} \subset T$, let $|\mathcal{A}|$ denote its cardinality. An admissible sequence is an increasing sequence of subsets $\{\mathcal{A}_n, n \geq 0\}$ of T, such that $|\mathcal{A}_0| = 1$ and for $n \geq 1$, $|\mathcal{A}_n| = 2^{2^n}$. Given $\alpha > 0$, we define the γ_α-functional as*

$$\gamma_\alpha(T, d) = \inf \sup_{t \in T} \sum_{n=0}^{\infty} \mathrm{Diam}(A_n(t)),$$

where $A_n(t)$ is the unique element of $\mathcal{A}_n$ that contains t, $\mathrm{Diam}(A_n(t))$ is the diameter of A_n according to d, and the infimum is over all admissible sequences of T.

A.1 Identifiability issues

It is was noted that our theory does not require the identifiability for each component in (2.2). For completeness, we discuss the following identifiable problems. To begin with, we state the uniqueness of the representation (2.2), which means that (2.2) is the only possible combination of the coefficients and functions under the minimal R components. There are three complications that result in the indeterminacy, where two of them are similar to that of CP decomposition. The first is about permutation and scaling, i.e.,

1. Permutation and scaling. Permutation means that the summation of CP components can be permuted, i.e.,

$$m(\mathbf{X}) = \nu + \frac{1}{s} \sum_{r \in \{1,\ldots,R\}} \langle \boldsymbol{\beta}_{r1} \circ \boldsymbol{\beta}_{r2} \circ \cdots \circ \boldsymbol{\beta}_{rD}, F_r(\mathbf{X}) \rangle ,$$

   while scaling means that for any constant $C \neq 0$,

$$\left\langle C\boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, \frac{1}{C} F_r(\mathbf{X}) \right\rangle = \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle ,$$

   where the scale C can also shift among $\{\boldsymbol{\beta}_{r,d}\}_{d=1}^{D}$.

The second is another possible combination of functions and the corresponding coefficients that can also represent $m(\mathbf{X})$ in (2.2), with the exception of permutation and scaling, i.e.,

2. Another possible combination. $m(\mathbf{X})$ can also be represented by

$$m(\mathbf{X}) = \nu + \frac{1}{s} \sum_{r=1}^{R} \left\langle \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D}, \bar{F}_r(\mathbf{X}) \right\rangle .$$

   This other combination is possible. For example, let

$$\bar{F}_1(\mathbf{X}) = \ldots = \bar{F}_R(\mathbf{X}) = F_1(\mathbf{X}) = \ldots = F_R(\mathbf{X}),$$

and

$$\mathbf{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}$$

Due to the non-uniqueness of CP decomposition of a tensor with rank R in general (Kolda and Bader, 2009), there is another rank decomposition for some $\mathbf{B}$ (see, e.g., Stegeman and Sidiropoulos, 2007), which will lead another combination to represent $m(\mathbf{X})$.

Besides, the constant shift also brings the indeterminacy.

3. Constant shift. For a constant C and a tensor $\mathbf{J} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ of which all the entries are 1,

$$\langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) - C\mathbf{J} \rangle = \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle + C',$$

where C' is a constant that can shift to the intercept ν of the model (2.2).

To avoid constant shift, we let $\int_0^1 f_r(x)\mathrm{d}x = 0$. This setting will not affect the expressive ability of the model (2.2). Now, we define the identifiability rigorously.

Definition 4 (Identifiability)**.** *Suppose $f_r \in \mathcal{F}$, where $\mathcal{F} = \{f : \int_0^1 f(x)\mathrm{d}x = 0, f \in \mathcal{C}([0,1])\}$, $r = 1, \ldots, R$ and $\{f_r\}_{r=1}^R$ is the minimal representation to make* (2.2) *hold. The minimal representation means that there does not exist one of the following two representations for $m(\mathbf{X})$, i.e.,*

i. $$m(\mathbf{X}) = \bar{\nu} + \tfrac{1}{s} \textstyle\sum_{r=1}^{\bar{R}} \left\langle \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D}, \bar{F}_r(\mathbf{X}) \right\rangle,$$

where $\bar{\nu} \in \mathbb{R}$, $\bar{\boldsymbol{\beta}}_{r,d} \in \mathbb{R}^{p_d \times \bar{R}}$, $(\bar{F}_r(\mathbf{X}))_{i_1,\ldots,i_D} = \bar{f}_r(\mathbf{X}_{i_1,\ldots,i_D}) \in \mathcal{F}$ and $\bar{R} < R$, or

ii. $$m(\mathbf{X}) = \tilde{\nu} + \tfrac{1}{s} \textstyle\sum_{r=1}^{R} \left\langle \tilde{\boldsymbol{\beta}}_{r,1} \circ \tilde{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \tilde{\boldsymbol{\beta}}_{r,D}, \tilde{F}_r(\mathbf{X}) \right\rangle,$$

where $\tilde{\nu} \in \mathbb{R}$, $\tilde{\boldsymbol{\beta}}_{r,d} \in \mathbb{R}^{p_d \times \bar{R}}$, $(\tilde{F}_r(\mathbf{X}))_{i_1,\ldots,i_D} = \tilde{f}_r(\mathbf{X}_{i_1,\ldots,i_D}) \in \mathcal{F}$ and Span$\{\tilde{f}_r\}_{r=1}^R \subsetneq$ Span$\{f_r\}_{r=1}^R$. We say the representation is identifiable if the components are unique up to permutation and scal-

ing. To be more specific if

$$
\begin{aligned}
m(\mathbf{X}) &= \nu + \frac{1}{s}\sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle \\
&= \bar{\nu} + \frac{1}{s}\sum_{r=1}^{R} \left\langle \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D}, \bar{F}_r(\mathbf{X}) \right\rangle,
\end{aligned}
$$

then $\nu = \bar{\nu}$*, and* $\{(\boldsymbol{\beta}_{r,1}, \boldsymbol{\beta}_{r,2}, \cdots, \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}))\}_{r=1}^{R}$ *and* $\{(\bar{\boldsymbol{\beta}}_{r,1}, \bar{\boldsymbol{\beta}}_{r,2}, \cdots, \bar{\boldsymbol{\beta}}_{r,D}, \bar{F}_r(\mathbf{X}))\}_{r=1}^{R}$ *are the same up to scaling.*

So far, we have demonstrated the identifiability issues and given the definition of identifiability with respect to the representation (2.2). We then list some sufficient conditions to achieve the identifiability, based on the fundamental idea of the identifiability for CP decomposition. Denote

$$
\mathbf{B}_d = (\boldsymbol{\beta}_{1,d}, \ldots, \boldsymbol{\beta}_{R,d}) \quad \text{for} \quad d = 1, \ldots, D,
$$

and k_{B_d} the k-rank of $\mathbf{B}_d$, which is defined as the maximum value k such that any k columns are linearly independent (Kruskal, 1977; Harshman, 1984). Then the following conditions in the two cases are sufficient to achieve the identifiability.

Case 1. Require that $\{f_r(x)\}_{r=1}^{R}$ is linearly independent.

i. If $\sum_{d=1}^{D} k_{B_d} \geq R + D$, then the decomposition (2.2) is unique up to permutation and scaling.

ii. If $D = 2$ and $R(R-1) \leq p_1(p_1-1)p_2(p_2-1)/2$, then the decomposition (2.2) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measure zero.

iii. If $D = 3$ and $R(R-1) \leq p_1p_2p_3(3p_1p_2p_3) - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, then the decomposition (2.2) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measures zero.

Case 2. Not require that $\{f_r(x)\}_{r=1}^{R}$ is linearly independent.

iv. (General) If $\sum_{d=1}^{D} k_{B_d} \geq 2R + D - 1$, then the decomposition (2.2) is unique up to permutation and scaling.

For simplicity, we present the general condition in the following theorem. In the proof of Theorem 3, we in fact prove all the aforementioned sufficient conditions.

Theorem 3. *(Identifiability) If*

$$\sum_{d=1}^{D} k_{B_d} \geq 2R + D - 1, \tag{A.1}$$

then the representation (2.2) *is unique up to permutation and scaling.*

Proof of Theorem 3

Proof. Suppose there is another representation of (2.2), i.e,

$$\begin{aligned} m(\mathbf{X}) &= \nu + \frac{1}{s}\sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle \\ &= \bar{\nu} + \frac{1}{s}\sum_{r=1}^{R} \left\langle \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D}, \bar{F}_r(\mathbf{X}) \right\rangle, \end{aligned} \tag{A.2}$$

where

$$(F_r(\mathbf{X}))_{i_1 i_2 \cdots i_D} = f_r(\mathbf{X}_{i_1 i_2 \cdots i_D}) \quad \text{and} \quad (\bar{F}_r(\mathbf{X}))_{i_1 i_2 \cdots i_D} = \bar{f}_r(\mathbf{X}_{i_1 i_2 \cdots i_D}),$$

with $f_r, \bar{f}_r \in \mathcal{F}$, $r = 1, \ldots, R$. We will show $\bar{\nu} = \nu$, as well as $\boldsymbol{\beta}_{r,d}$ and $\bar{\boldsymbol{\beta}}_{r,d}$, f_r and $\bar{f}_r$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, are the same up to permutation and scaling under some conditions, respectively.

Using the definition of $\mathbf{F}$, such as $\int_0^1 f(x)\mathrm{d}x = 0$ for $f \in \mathcal{F}$, we can obtain $\nu = \bar{\nu}$ by integration over the domain of $\mathbf{X}$ in (A.2). In the remaining sum of inner products, we consider the following arguments. Suppose the minimal bases of the vector space

$$\mathrm{Span}\{f_r(x), r = 1, \ldots, R\} \quad \text{and} \quad \mathrm{Span}\{\bar{f}_r(x), r = 1, \ldots, R\}$$

are $\{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star}$ and $\{\bar{\psi}_{\bar{k}^\star}(x)\}_{\bar{k}^\star=1}^{\bar{K}^\star}$, respectively. In other words, each f_r and $\bar{f}_r$ can be written in

a unique way as a linear combination of $\{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star}$ and $\{\bar{\psi}_{\bar{k}^\star}(x)\}_{\bar{k}^\star=1}^{\bar{K}^\star}$, respectively. To be more specific,

$$f_r(x) = \sum_{k^\star=1}^{K^\star} \eta_{r,k^\star}\psi_{k^\star}(x) \quad \text{and} \quad \bar{f}_r(x) = \sum_{\bar{k}^\star=1}^{\bar{K}^\star} \bar{\eta}_{r,\bar{k}^\star}\bar{\psi}_{\bar{k}^\star}(x).$$

For notational convenience, we let $\Psi(\mathbf{X})_{\boldsymbol{j},k^\star} = \psi_{k^\star}(\mathbf{X}_{\boldsymbol{j}})$, $k^\star = 1,\ldots,K^\star$ and $\bar{\Psi}(\mathbf{X})_{\boldsymbol{j},\bar{k}^\star} = \bar{\psi}_{\bar{k}^\star}(\mathbf{X}_{\boldsymbol{j}})$, $\bar{k}^\star = 1,\ldots,\bar{K}^\star$, where $\boldsymbol{j} \in \mathcal{J}$. We also denote

$$\mathbf{A}^f = \frac{1}{s}\sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \boldsymbol{\eta}_r, \tag{A.3}$$

and

$$\bar{\mathbf{A}}^f = \frac{1}{s}\sum_{r=1}^{R} \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D} \circ \bar{\boldsymbol{\eta}}_r, \tag{A.4}$$

where $\boldsymbol{\eta}_r = (\eta_{r,1},\cdots,\eta_{r,K})^\intercal$ and $\bar{\boldsymbol{\eta}}_r = (\bar{\eta}_{r,1},\cdots,\bar{\eta}_{r,K})^\intercal$, for $r = 1,\ldots,R$. Since we have shown $\nu = \bar{\nu}$ in the previous arguments, it is trivial to see that the remaining summation of CP components in (A.2) equals, i.e.,

$$\big\langle \mathbf{A}^f, \Psi(\mathbf{X}) \big\rangle = \big\langle \bar{\mathbf{A}}^f, \bar{\Psi}(\mathbf{X}) \big\rangle. \tag{A.5}$$

The rest of proof includes three steps. At first, we will show

$$\mathrm{Span}\{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star} = \mathrm{Span}\{\bar{\psi}_{\bar{k}^\star}(x)\}_{\bar{k}^\star=1}^{\bar{K}^\star}. \tag{A.6}$$

Based on (A.6), we can chose $\{\bar{\psi}_{\bar{k}^\star}(x)\}_{\bar{k}^\star=1}^{\bar{K}^\star} = \{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star}$ and rewrite (A.5) as

$$\big\langle \mathbf{A}^f, \Psi(\mathbf{X}) \big\rangle = \big\langle \bar{\mathbf{A}}^f, \Psi(\mathbf{X}) \big\rangle. \tag{A.7}$$

Secondly, we will show $\mathbf{A}^f = \bar{\mathbf{A}}^f$ in (A.7). In the end, we will take the advantages of identifiable theory about CP decomposition and complete the proof.

To show (A.6), we assume there exists k_0 such that $\bar{\psi}_{k_0}(x)$ is linearly independent of $\{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star}$. For each $\boldsymbol{j} \in \mathcal{J}$, we take integration for other predictors over their domain, then by Lemma 1, we

get

$$\sum_{k^\star=1}^{K^\star} A^f_{\boldsymbol{j},k^\star}\psi_{k^\star}(X_{\boldsymbol{j}}) - \sum_{\bar{k}^\star\neq k_0}^{\bar{K}^\star} \bar{A}^f_{\boldsymbol{j},\bar{k}^\star}\bar{\psi}_{\bar{k}^\star}(X_{\boldsymbol{j}}) - \bar{A}^f_{\boldsymbol{j},k_0}\bar{\psi}_{k_0}(X_{\boldsymbol{j}}) = 0,$$

for $X_{\boldsymbol{j}} \in [0,1]$. Note that $\bar{\psi}_{k_0}(x)$ is independent of $\{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star}$ and $\{\bar{\psi}_{\bar{k}^\star}(x)\}_{i\neq k_0}$, then $\bar{A}^f_{\boldsymbol{j},k_0} = 0$, for $\boldsymbol{j} \in \mathcal{J}$. Assume there exists r_0 such that $\bar{\eta}_{r_0,k_0} \neq 0$, then there exists $\{\tilde{f}_r\}_{r=1}^R$, where $\tilde{f}_r(x) = \sum_{k^\star\neq k_0} \bar{\eta}_{r,k^\star}\bar{\psi}_i(x)$ and Span$\{\tilde{f}_r\}_{r=1}^R \subsetneq$ Span$\{f_r\}_{r=1}^R$, such the representation (2.2) holds. This does not agree with the minimal representation assumption. As a result, $\bar{\eta}_{r,k_0} = 0$ for $r = 1,\ldots,R$, then $\{\bar{f}_r(x)\}_{r=1}^R$ can be represented by $\{\bar{\psi}_{k^\star}(x)\}_{k^\star\neq k_0}$, which leads a contradiction to that $\{\bar{\psi}_{\bar{k}^\star}(x)\}_{\bar{k}^\star=1}^{\bar{K}^\star}$ is a minimal basis. Therefore (A.6) holds and $\bar{K}^\star = K^\star$.

To show $\mathbf{A}^f = \bar{\mathbf{A}}^f$ in (A.7), we let $\mathbf{A}^{f,\star} = \mathbf{A}^f - \bar{\mathbf{A}}^f$. It is implies that

$$\left\langle \mathbf{A}^{f,\star}, \Psi(\mathbf{X}) \right\rangle = 0,$$

for all $\mathbf{X}$. Assuming $\mathbf{A}^{f,\star} \neq \mathbf{0}$, there exists $\boldsymbol{j}_0 \in \mathcal{J}$ such that $(A^{f,\star}_{\boldsymbol{j}_0,1},\ldots,A^{f,\star}_{\boldsymbol{j}_0,K^\star}) \neq \mathbf{0}$. We fix $\{X_{\boldsymbol{j}}\}_{\boldsymbol{j}\neq\boldsymbol{j}_0}$ at some values and let the corresponding value $C_{-\boldsymbol{j}_0} = \sum_{\boldsymbol{j}\neq\boldsymbol{j}_0}\sum_{k^\star=1}^{K^\star} A^{f,\star}_{\boldsymbol{j},k^\star} f_{k^\star}(X_{\boldsymbol{j}})$. Then

$$\sum_{k^\star=1}^{K^\star} A^{f,\star}_{\boldsymbol{j}_0,k^\star}\psi_{k^\star}(X_{\boldsymbol{j}_0}) + C_{-\boldsymbol{j}_0} = 0, \tag{A.8}$$

for $X_{\boldsymbol{j}_0} \in [0,1]$. By integration over $X_{\boldsymbol{j}_0}$ on both sides, we obtain

$$\sum_{k^\star=1}^{K^\star} A^{f,\star}_{\boldsymbol{j}_0,i^\star} w_{k^\star} + C_{-\boldsymbol{j}_0} = 0,$$

where $w_{k^\star} = \int_0^1 \psi_{k^\star}(x)\mathrm{d}x$, $k^\star = 1,\ldots,K^\star$. By Lemma 1, $\sum_{k^\star=1}^{K^\star} A^{f,\star}_{\boldsymbol{j}_0,k^\star} w_{k^\star} = 0$, which implies $C_{-\boldsymbol{j}_0} = 0$. Combining the independence and (A.8) yields $A^{f,\star}_{\boldsymbol{j}_0,k^\star} = 0$ for $k^\star = 1,\ldots,K^\star$. Thus $\mathbf{A}^{f,\star} = \mathbf{0}$ and we have $\mathbf{A}^f = \bar{\mathbf{A}}^f$.

Since R is the minimal, (A.3) is a rank decomposition of $\mathbf{A}^f$. We can claim that if the rank decomposition of $\mathbf{A}^f$ is unique up to permutation and scaling, then the representation (2.2) is unique up to scaling and permutation. To see this, we can assume the rank decomposition of $\mathbf{A}^f$

is unique up to permutation and scaling. Thus the decomposition (A.4) and the decomposition (A.3) are the same up to permutation and scaling. Therefore the representation (2.2) is unique up to permutation and scaling. Now, to make the representation (2.2) unique up to permutation and scaling, we can use the common arguments about the uniqueness of rank decomposition. Recall that $\mathbf{B}_d = (\boldsymbol{\beta}_{1,d}, \ldots, \boldsymbol{\beta}_{R,d})$, $d = 1, \ldots, D$ and the k-rank of a matrix $\mathbf{B}_d$, denoted as k_{B_d}, is defined as the maximum value k such that any k columns are linearly independent. For convenience, we let $\mathbf{B}_{D+1} := \boldsymbol{\eta} = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_R)$ and let $k_{B_{D+1}}$ be its k-rank. To make the CP decomposition of $\mathbf{A}^f$ unique, we have the following sufficient conditions

1. (General) (Sidiropoulos and Bro, 2000) The decomposition (A.3) is unique up to permutation and scaling if $\sum_{d=1}^{D+1} k_{B_d} \geq 2R + D$.

2. (De Lathauwer, 2006) When $D + 1 = 3$, $R \leq K$ and $R(R-1) \leq p_1(p_1-1)p_2(p_2-1)/2$, the decomposition (A.3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measure zero.

3. (De Lathauwer, 2006) When $D+1 = 4$, $R \leq K$ and $R(R-1) \leq p_1p_2p_3(3p_1p_2p_3) - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, the decomposition (A.3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measures zero.

Now we consider two cases, i.e,

Case 1. If $\{f_r(x)\}_{r=1}^R$ is linearly independent, then $k_{B_{D+1}} = R$. We have the following sufficient conditions.

i. If $\sum_{d=1}^{D} k_{B_d} \geq R+D$, then the decomposition (2.2) is unique up to permutation and scaling.

ii. If $D = 2$ and $R(R-1) \leq p_1(p_1-1)p_2(p_2-1)/2$, then the decomposition (2.2) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measure zero.

iii. If $D = 3$ and $R(R-1) \leq p_1p_2p_3(3p_1p_2p_3) - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, then the decomposition (2.2) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measures zero.

Case 2. If we do not know whether $\{f_r(x)\}_{r=1}^R$ is linearly independent or not, we can also use the fact that $k_{B_{D+1}} \geq 1$, which yields the following general sufficient condition.

iv. (General) If $\sum_{d=1}^D k_{B_d} \geq 2R + D - 1$, then the decomposition (2.2) is unique up to permutation and scaling.

Since $\{f_r\}_{r=1}^R$ are allowed to be the same in the model, we can use the forth sufficient condition, i.e.,

$$\sum_{d=1}^{D} k_{B_d} \geq 2R + D - 1,$$

which is also used as a condition to make the tensor linear model identifiable (Zhou et al., 2013).

□

A.2 Estimation

A.2.1 Equivalent basis

To begin with, we define some notations which will be used later. Suppose $\{b'_k(x)\}_{k=1}^K$ is the truncated power basis and $\{b_k(x)\}_{k=1}^K$ is the B-spline basis. Let $u_k = \int_0^1 b_k(x)\mathrm{d}x$ and $u'_k = \int_0^1 b'_k(x)\mathrm{d}x$. Denote $\Phi(\mathbf{X}), \Phi'(\mathbf{X}) \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D \times K}$ be the tensor formed from the bases, which means $(\Phi(\mathbf{X}))_{\boldsymbol{j},k} = b_k(X_{\boldsymbol{j}})$ and $(\Phi'(\mathbf{X}))_{\boldsymbol{j},k} = b'_k(X_{\boldsymbol{j}})$, $\boldsymbol{j} \in \mathcal{J}$, $k = 1, \ldots, K$. We define two function classes,

$$\mathcal{M}_1 = \left\{ m(\mathbf{X}) : m(\mathbf{X}) = \nu_1 + \frac{1}{s}\sum_{r=1}^{R} \langle \boldsymbol{\beta}_{1r,1} \circ \boldsymbol{\beta}_{1r,2} \circ \cdots \circ \boldsymbol{\beta}_{1r,D} \circ \boldsymbol{\alpha}_{1r}, \Phi(\mathbf{X}) \rangle, \sum_{k=1}^{K} \alpha_{1r,k} u_k = 0 \right\},$$

and

$$\mathcal{M}_2 = \left\{ m(\mathbf{X}) : m(\mathbf{X}) = \nu_2 + \frac{1}{s}\sum_{r=1}^{R} \langle \boldsymbol{\beta}_{2r,1} \circ \boldsymbol{\beta}_{2r,2} \circ \cdots \circ \boldsymbol{\beta}_{2r,D} \circ \boldsymbol{\alpha}_{2r}, \Phi'(\mathbf{X}) \rangle, \sum_{k=1}^{K} \alpha_{2r,k} u'_k = 0 \right\},$$

where $\nu_l \in \mathbb{R}$, $\boldsymbol{\beta}_{lr,d} \in \mathbb{R}^{p_d}$ and $\boldsymbol{\alpha}_{lr} = (\alpha_{lr,1}, \cdots, \alpha_{lr,K})^\intercal \in \mathbb{R}^K$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, $l = 1, 2$. Particularly, one basis function of the truncated power basis is a constant 1. Without loss of generality, we let $b_1'(x) = 1$ and denote the tensor $\tilde{\Phi}(\mathbf{X}) \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K-1}$ formed by the remaining basis functions, which means $(\tilde{\Phi}(\mathbf{X}))_{i_1,\ldots,i_D,k} = b_{k+1}'(X_{i_1 \cdots i_D})$, $\boldsymbol{j} \in \mathcal{J}$, $k = 1, \ldots, K-1$. We define the following function class that is removed the linear constraints, i.e.,

$$\mathcal{M}_3 = \left\{ m(\mathbf{X}) : m(\mathbf{X}) = \nu + \frac{1}{s} \sum_{r=1}^{R} \left\langle \boldsymbol{\beta}_{3r,1} \circ \boldsymbol{\beta}_{3r,2} \circ \cdots \circ \boldsymbol{\beta}_{3r,D} \circ \boldsymbol{\alpha}_{3r}, \tilde{\Phi}(\mathbf{X}) \right\rangle \right\},$$

where $\nu_3 \in \mathbb{R}$, $\boldsymbol{\beta}_{3r,d} \in \mathbb{R}^{p_d}$, and $\boldsymbol{\alpha}_{3r} = (\alpha_{3r,1}, \cdots, \alpha_{3r,K-1})^\intercal \in \mathbb{R}^{K-1}$, $r = 1, \ldots, R$, $d = 1, \ldots, D$.

By the following Theorem 4, we can remove the linear constraints in (3.3) and use any equivalent spline basis to develop our theory.

Theorem 4. $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}_3$.

Proof. Firstly, we will prove $\mathcal{M}_1 = \mathcal{M}_2$. For each $m_1(\mathbf{X}) \in \mathcal{M}_1$. By the property of spline basis (see, e.g., Chapter 3 of Ruppert et al. (2003)), there exists an invertible matrix $\mathbf{Q}$ such that $\mathbf{b}(x) = \mathbf{Q}\mathbf{b}'(x)$. It is straightforward to see $\mathcal{M}_1 = \mathcal{M}_2$.

Secondly, we will prove $\mathcal{M}_2 \subset \mathcal{M}_3 \subset \mathcal{M}_2$. For notational simplicity, denote

$$\mathbf{B}_{lr} = \boldsymbol{\beta}_{lr,1} \circ \ldots \circ \boldsymbol{\beta}_{lr,D}, \quad \text{for } l = 2, 3,$$

For each $m_2(\mathbf{X}) \in \mathcal{M}_2$, take $\mathbf{B}_{3r} = \mathbf{B}_{2r}$, $v_3 = v_2 + 1/s \sum_{r=1}^{R} \langle \mathbf{B}_{2r}, \alpha_{2r,1}\mathbf{J} \rangle$ and $\alpha_{3r,k} = \alpha_{2r,k+1}$, for $k = 1, \ldots, K-1$. Then we have $m_2(\mathbf{X}) = m_3(\mathbf{X}) \in \mathcal{M}_3$ and $\mathcal{M}_2 \subset \mathcal{M}_3$. For each $m_3(\mathbf{X}) \in \mathcal{M}_3$. Suppose $\sum_{k=1}^{K-1} \alpha_{3r,k} u_{k+1}' = C_r$, it is trivial to see $u_1' \neq 0$. We can choose $\alpha_{1r,2} = -C_r/u_1'$, $\alpha_{2r,k+1} = \alpha_{3r,k}$ for $k = 1, \ldots, K-1$ so that $\boldsymbol{\alpha}_{2r}$ satisfies the constraint in $\mathcal{M}_2$. Taking $\nu_2 = \nu_3 + \sum_{r=1}^{R} \langle \mathbf{B}_{3r}, C_r/u_1'\mathbf{J} \rangle$, $\mathbf{B}_{2r} = \mathbf{B}_{3r}$, it is trivial to see $m_3(\mathbf{X}) = m_2(\mathbf{X}) \in \mathcal{M}_2$. Thus $\mathcal{M}_3 \subset \mathcal{M}_2$ and we get $\mathcal{M}_3 = \mathcal{M}_2$.

□

A.2.2 Rescaling strategy for the elastic net

For the elastic net penalty, denote

$$G(\mathbf{B}_1, \ldots, \mathbf{B}_D) = \lambda_1 \sum_{r=1}^{R} G_r(\{\boldsymbol{\beta}_{r,d}\}_d, \lambda_2),$$

where

$$G_r(\{\boldsymbol{\beta}_{r,d}\}_d, \lambda_2) = \sum_{d=1}^{D} \frac{1}{2}(1-\lambda_2)\|\boldsymbol{\beta}_{r,d}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}_{r,d}\|_1.$$

Let $\tilde{\rho}_{r,d} = \log \rho_{r,d}$, then the above optimization problem (3.6) becomes

$$\begin{aligned} &\underset{\tilde{\rho}_{r,1},\cdots,\tilde{\rho}_{r,D}}{\arg\min} \sum_{d=1}^{D} \frac{1}{2}(1-\lambda_2)\|\boldsymbol{\beta}_{r,d}\|_2^2 \exp^2\{\tilde{\rho}_{r,d}\} + \lambda_2 \|\boldsymbol{\beta}_{r,d}\|_1 \exp\{\tilde{\rho}_{r,d}\} \\ &\text{s.t.} \quad \sum_{d=1}^{D} \tilde{\rho}_{r,d} = 0, \end{aligned} \tag{A.9}$$

which is a convex problem. Using the Lagrangian method and Newton's method, we can get the solution.

A.2.3 Proof of Proposition 1

Proof. Suppose $\boldsymbol{\theta}^{\boldsymbol{\rho}} = (\tilde{\nu}, \mathbf{B}_1^{\boldsymbol{\rho}}, \ldots, \mathbf{B}_D^{\boldsymbol{\rho}}, \tilde{\mathbf{B}}_{D+1}) \in \Theta(\boldsymbol{\theta})$, then there exists $\{\rho_{r,d}\}_{r,d}$ satisfying $\prod_d \rho_{r,d} = 1$ for $r = 1, \ldots, R$ such that $\mathbf{B}_d^{\boldsymbol{\rho}} = (\rho_{1,d}\boldsymbol{\beta}_{1,d}, \ldots, \rho_{R,d}\boldsymbol{\beta}_{R,d})$ for $d = 1, \ldots, D$. By definition, for each $r = 1, \ldots, R$

$$G_r(\{\hat{\rho}_{r,d}\boldsymbol{\beta}_{r,d}\}_d, \lambda_2) \leq G_r(\{\rho_{r,d}\boldsymbol{\beta}_{r,d}\}_d, \lambda_2),$$

thus

$$LG(\bar{\boldsymbol{\theta}}) \leq LG(\boldsymbol{\theta}^{\boldsymbol{\rho}}).$$

Note that (A.9) is a strictly convex problem if $\boldsymbol{\beta}_{r,d} \neq \mathbf{0}$ for $r = 1, \ldots, R, d = 1, \ldots, D$. Thus $\bar{\boldsymbol{\theta}}$ is the unique minimizer in $\Theta(\boldsymbol{\theta})$ and we complete the proof. □

A.2.4 Proof of Proposition 2

Proof. We first note that with our rescaling strategy, the objective function is non-increase after each iteration in our algorithm. Using the same arguments of Proposition 1 of Zhou et al. (2013), we can get the desired result. □

A.3 Proof of Theorem 1

Proof. Since $\hat{A}$ and $\hat{\nu}$ is a solution of (3.3), we have

$$\sum_{i=1}^{n}\left(y_i-\hat{\nu}-\frac{1}{s}\langle\hat{\mathbf{A}},\Phi(\mathbf{X}_i)\rangle\right)^2\leq\sum_{i=1}^{n}\left(y_i-\nu_0-\frac{1}{s}\langle\mathbf{A}_0,\Phi(\mathbf{X}_i)\rangle\right)^2.$$

Using the definition of $\mathcal{I}$ in (4.1), the aforementioned inequality is equivalent to

$$\sum_{i=1}^{n}\left(y_i-\frac{1}{s}\langle\hat{\mathbf{A}}^\flat,\Phi(\mathbf{X}_i)\rangle\right)^2\leq\sum_{i=1}^{n}\left(y_i-\frac{1}{s}\langle\mathbf{A}_0^\flat,\Phi(\mathbf{X}_i)\rangle\right)^2. \tag{A.10}$$

Let $\mathbf{A}^\sharp=\hat{\mathbf{A}}^\flat-\mathbf{A}_0^\flat$, $\mathbf{a}^\sharp=\mathrm{vec}(\mathbf{A}^\sharp)$, $\mathbf{a}_0^\flat=\mathrm{vec}(\mathbf{A}_0^\flat)$ and $\mathbf{Z}=(\mathbf{z}_1,\ldots,\mathbf{z}_n)^\intercal\in\mathbb{R}^{n\times s}$, where $\mathbf{z}_i=\mathrm{vec}\{\Phi(\mathbf{X}_i)\}$, $i=1,\ldots,n$. In fact, $\mathbf{Z}$ can be regarded as the "design" matrix formed by the spline basis. Using (A.10) and working out the squares, we obtain

$$\frac{1}{s^2}\|\mathbf{Z}\mathbf{a}^\sharp\|_2^2\leq 2\left\langle\frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp,\boldsymbol{\epsilon}\right\rangle+2\left\langle\frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp,\mathbf{y}-\boldsymbol{\epsilon}-\frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat\right\rangle, \tag{A.11}$$

where $\mathbf{y}=(y_1,\cdots,y_n)^\intercal$. By Lemma 1, we have $\sum_{k=1}^{K}A^\sharp_{\boldsymbol{j},k}u_k=0$ for $\boldsymbol{j}\in\mathcal{J}/\{(1,\cdots,1)\}$, where

$$u_k=\int_0^1 b_k(x)\mathrm{d}x.$$

Since $\mathrm{rank}(\mathbf{A}_0^\flat)\leq R_0+1$, $\mathrm{rank}(\hat{\mathbf{A}}^\flat)\leq R+1$, it is trivial to see $\mathrm{rank}(\mathbf{A}^\sharp)\leq R_0+R+2$. To finish the proof, we will find the upper bound of the right hand side and the lower bound of the left hand side with respect to $\|\mathbf{a}^\sharp\|_2$ in $(A.11)$.

Firstly, we will find the upper bound of $\langle \mathbf{Z}\mathbf{a}^\sharp, \boldsymbol{\epsilon}\rangle$. To simplify the notations, let

$$\mathcal{P} = \left\{ \frac{\mathrm{vec}(\mathbf{A})}{\|\mathbf{A}\|_{HS}} : \sum_{k=1}^{K} A_{\boldsymbol{j},k} u_k = 0, \text{ for } \boldsymbol{j} \in \mathcal{J}/\{(1,\ldots,1)\}, \mathrm{rank}(\mathbf{A}) \le R_1 \right\},$$

where $R_1 = R + R_0 + 2 \le 2R+2$. By (A.39) of Lemma 4, if $n > C\tilde{h}_n^2 h_n^{-2} w^2(\mathcal{P})$ for some $C > 0$,

$$C_1 n h_n \|\mathbf{a}^\sharp\|_2^2 \le \|\mathbf{Z}\mathbf{a}^\sharp\|_2^2 \le C_2 n h_n \|\mathbf{a}^\sharp\|_2^2 \tag{A.12}$$

with probability as least $1 - 2\exp\{-C_3 w^2(\mathcal{P})\}$. By Lemma 3, the Gaussian width $w(\mathcal{P}) \le C_4(R^{D+1} + R\sum_{i=1}^D p_i + RK)^{1/2}$. In the following part, we assume $n > C\tilde{h}_n^2 h_n^{-2}(R^{D+1} + R\sum_{i=1}^D p_i + RK)$ for some $C > 0$, then (A.12) holds. By Lemma 5 and (A.12), we have the following upper bound

$$\langle \mathbf{Z}\mathbf{a}^\sharp, \boldsymbol{\epsilon}\rangle \le \|\mathbf{a}^\sharp\|_2 \, \mathcal{O}_{\mathrm{p}}\left(\left\{ n h_n \left(R^{D+1} + \sum_{i=1}^{D} R p_i + RK \right) \right\}^{1/2} \right). \tag{A.13}$$

Secondly, we find the upper bound of $\langle \mathbf{Z}\mathbf{a}^\sharp, \mathbf{y} - \boldsymbol{\epsilon} - \mathbf{Z}\mathbf{a}_0^\flat \rangle$. Note that

$$\begin{aligned}
\left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\|_2^2 &= \sum_{i=1}^{n} \left| \frac{1}{s} \sum_{r=1}^{R_0} \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i)\rangle - \langle \mathbf{A}_0, \Phi(\mathbf{X}_i)\rangle \right|^2 \\
&\le \sum_{i=1}^{n} \left(\frac{1}{s} \sum_{r=1}^{R_0} \left| \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i)\rangle - \langle \mathbf{B}_{0r} \circ \boldsymbol{\alpha}_{0r}, \Phi(\mathbf{X}_i)\rangle \right| \right)^2 \\
&\le \sum_{i=1}^{n} \left\{ \frac{1}{s} \sum_{r=1}^{R_0} \frac{C}{K^\tau} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1 \right\}^2 \\
&= \mathcal{O}_{\mathrm{p}}\left(\left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\}^2 \frac{n}{K^{2\tau}} \right).
\end{aligned}$$

Using the Cauchy-Schwarz inequality and (A.12), it is shown that

$$\begin{aligned}
&\left\langle \frac{1}{s}\mathbf{Z}\mathbf{a}^{\sharp}, \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^{\flat} \right\rangle \\
&\leq \left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^{\flat} \right\|_2 \left\| \frac{1}{s}\mathbf{Z}\mathbf{a}^{\sharp} \right\|_2 \\
&= \frac{1}{s}\|\mathbf{Z}\mathbf{a}^{\sharp}\|_2 \, \mathcal{O}_{\mathrm{p}}\left(\left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{\sqrt{n}}{K^{\tau}} \right) \\
&= \frac{1}{s}\|\mathbf{a}^{\sharp}\|_2 \, \mathcal{O}_{\mathrm{p}}\left(\left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{n\sqrt{h_n}}{K^{\tau}} \right).
\end{aligned} \tag{A.14}$$

Finally, plugging (A.13) and (A.14) into (A.11), we get

$$\begin{aligned}
\frac{1}{s^2}\|\mathbf{Z}\mathbf{a}^{\sharp}\|_2^2 \leq \frac{1}{s}\|\mathbf{a}^{\sharp}\|_2 \, \mathcal{O}_{\mathrm{p}}\left(\left\{ nh_n \left(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK \right) \right\}^{\frac{1}{2}} \right) \\
+ \mathcal{O}_{\mathrm{p}}\left(\left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{n\sqrt{h_n}}{K^{\tau}} \right).
\end{aligned} \tag{A.15}$$

It follows from (A.12) and (A.15) that

$$\frac{1}{\sqrt{s}}\|\mathbf{a}^{\sharp}\|_2 = \mathcal{O}_{\mathrm{p}}\left(\left\{ \frac{sK(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK)}{n} \right\}^{\frac{1}{2}} + \left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{\sqrt{s}} \right\} \frac{1}{K^{\tau - 1/2}} \right),$$

which completes the proof of (4.5). Further, by Assumption 1 and (A.24) of Lemma 2, we have

$$\|\hat{m}(\mathbf{X}) - m(\mathbf{X})\|_{L_2}^2 \leq C_5 h_n \frac{1}{s^2}\|\hat{\mathbf{A}}^{\flat} - \mathbf{A}_0^{\flat}\|_{HS}^2 = C_5 h_n \frac{1}{s^2}\|\mathbf{a}^{\sharp}\|_2^2, \tag{A.16}$$

where C_5 is a constant, which will complete the proof of (4.6). □

A.4 Proof of Theorem 2

Proof. Since the arguments used in the proof of Theorem 1 have non-asymptotic versions, we can show the consistency of the penalized estimator, similarly. To simplify the notations, let

$$\hat{G} = \sum_{r=1}^{R} \sum_{d=1}^{D} \sum_{i=1}^{p_d} P_{\lambda}(\hat{\beta}_{r,di}).$$

Similar to the proof of Theorem 1, we can obtain

$$\sum_{i=1}^{n}\left(y_i-\hat{\nu}-\frac{1}{s}\langle\hat{\mathbf{A}},\Phi(\mathbf{X}_i)\rangle\right)^2+\hat{G}\leq\sum_{i=1}^{n}\left(y_i-\nu_0-\frac{1}{s}\langle\mathbf{A}_0,\Phi(\mathbf{X}_i)\rangle\right)^2+G_0.$$

Since $\hat{G}\geq 0$, we have

$$\sum_{i=1}^{n}\left(y_i-\frac{1}{s}\langle\hat{\mathbf{A}}^\flat,\Phi(\mathbf{X}_i)\rangle\right)^2\leq\sum_{i=1}^{n}\left(y_i-\frac{1}{s}\langle\mathbf{A}_0^\flat,\Phi(\mathbf{X}_i)\rangle\right)^2+G_0. \tag{A.17}$$

Let $\mathbf{A}^\sharp=\hat{\mathbf{A}}^\flat-\mathbf{A}_0^\flat$, $\mathbf{a}^\sharp=\text{vec}(\mathbf{A}^\sharp)$, $\mathbf{a}_0^\flat=\text{vec}(\mathbf{A}_0^\flat)$ and $\mathbf{Z}=(\mathbf{z}_1,\ldots,\mathbf{z}_n)^\intercal\in\mathbb{R}^{n\times s}$, where $\mathbf{z}_i=\text{vec}\{\Phi(\mathbf{X}_i)\}$, $i=1,\ldots,n$. In fact, $\mathbf{Z}$ can be regarded as the "design" matrix formed by the spline basis. Using (A.17) and working out the squares, we obtain

$$\frac{1}{s^2}\|\mathbf{Z}\mathbf{a}^\sharp\|_2^2\leq 2\left\langle\frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp,\boldsymbol{\epsilon}\right\rangle+2\left\langle\frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp,\mathbf{y}-\boldsymbol{\epsilon}-\frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat\right\rangle+G_0, \tag{A.18}$$

where $\mathbf{y}=(y_1,\cdots,y_n)^\intercal$. By Lemma 1 and Lemma 6, we have $\sum_{k=1}^{K}A^\sharp_{\boldsymbol{j},k}u_k=0$ for $\boldsymbol{j}\in\mathcal{J}/\{(1,\cdots,1)\}$, where

$$u_k=\int_0^1 b_k(x)\mathrm{d}x.$$

Since $\text{rank}(\mathbf{A}_0^\flat)\leq R_0+1$, $\text{rank}(\hat{\mathbf{A}}^\flat)\leq R+1$, it is trivial to see $\text{rank}(\mathbf{A}^\sharp)\leq R_0+R+2$. To finish the proof, we will try to find the upper bound of the right hand side and the lower bound of the left hand side with respect to $\|\mathbf{a}^\sharp\|_2$ in (A.18).

Firstly, we will find the upper bound of $\langle\mathbf{Z}\mathbf{a}^\sharp,\boldsymbol{\epsilon}\rangle$. To simplify the notation, let

$$\mathcal{P}=\left\{\frac{\text{vec}(\mathbf{A})}{\|\mathbf{A}\|_{HS}}:\sum_{k=1}^{K}A_{\boldsymbol{j},k}u_k=0,\ \text{for}\ \boldsymbol{j}\in\mathcal{J}/\{(1,\ldots,1)\},\text{rank}(\mathbf{A})\leq R_1\right\},$$

where $R_1=R+R_0+2\leq 2R+2$. Recall that if $n>C\tilde{h}_n^2h_n^{-2}w^2(\mathcal{P})$, for some constant C, then

$$C_1nh_n\|\mathbf{a}^\sharp\|_2^2\leq\|\mathbf{Z}\mathbf{a}^\sharp\|_2^2\leq C_2nh_n\|\mathbf{a}^\sharp\|_2^2 \tag{A.19}$$

with probability as least $1 - 2\exp\{-Cw^2(\mathcal{P})\}$, and the Gaussian width $w(\mathcal{P}) \leq C(R^{D+1} + R\sum_{i=1}^{D} p_i + RK)^{1/2}$. In the following part, we assume $n > C\tilde{h}_n^2 h_n^{-2}(R^{D+1} + R\sum_{i=1}^{D} p_i + RK)$ for some $C > 0$, then (A.19) holds. By Lemma 5, we have the following upper bound

$$\langle \mathbf{Z}\mathbf{a}^\sharp, \boldsymbol{\epsilon} \rangle \leq C\|\mathbf{a}^\sharp\|_2 \left\{ nh_n \left(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK \right) \right\}^{1/2}, \tag{A.20}$$

with probability at least

$$1 - C_3 \exp\left\{ - C_4 \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right) \right\}.$$

Secondly, we find the upper bound of $\langle \mathbf{Z}\mathbf{a}^\sharp, \mathbf{y} - \boldsymbol{\epsilon} - \mathbf{Z}\mathbf{a}_0^\flat \rangle$. Note that

$$\begin{aligned}
\left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\|_2^2 &= \sum_{i=1}^{n} \left| \frac{1}{s} \sum_{r=1}^{R_0} \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i) \rangle - \langle \mathbf{A}_0, \mathbf{B}(\mathbf{X}_i) \rangle \right|^2 \\
&\leq \sum_{i=1}^{n} \left(\frac{1}{s} \sum_{r=1}^{R_0} \left| \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i) \rangle - \langle \mathbf{B}_{0r} \circ \boldsymbol{\alpha}_{0r}, \mathbf{B}(\mathbf{X}_i) \rangle \right| \right)^2 \\
&\leq \sum_{i=1}^{n} \left\{ \frac{1}{s} \sum_{r=1}^{R_0} \frac{C}{K^\tau} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1 \right\}^2 \\
&= C \left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\}^2 \frac{n}{K^{2\tau}}.
\end{aligned}$$

It follows from the Cauchy-Swarchz inequality and (A.19) that

$$\begin{aligned}
&\left\langle \frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp, \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\rangle \\
&\leq \left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\|_2 \left\| \frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp \right\|_2 \\
&\leq \frac{C}{s}\|\mathbf{Z}\mathbf{a}^\sharp\|_2 \left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{\sqrt{n}}{K^\tau} \\
&\leq \frac{C}{s}\|\mathbf{a}^\sharp\|_2 \left\{ \frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{n\sqrt{h_n}}{K^\tau},
\end{aligned} \tag{A.21}$$

with probability at least $1 - 2\exp\{-Cw^2(\mathcal{P})\}$.

Thirdly, applying (A.19), (A.20) and (A.21) to (A.18), we get

$$\frac{1}{s^2}\|\mathbf{a}^\sharp\|_2^2 \leq \frac{\delta_3}{s}\|\mathbf{a}^\sharp\|_2 + \frac{1}{nh_n}G_0, \tag{A.22}$$

with probability at least

$$1 - C_5 \exp\left\{ - C_6\left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK\right)\right\}, \tag{A.23}$$

where

$$\delta_3 = C\left\{\frac{K\left(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK\right)}{n}\right\}^{\frac{1}{2}} + C\left\{\frac{\sum_{r=1}^{R_0}\|\text{vec}(\mathbf{B}_{0r})\|_1}{s}\right\}\frac{1}{K^{\tau-1/2}}.$$

By solving the second order inequality (A.22), we obtain

$$\frac{1}{s}\|\mathbf{a}^\sharp\|_2 \leq \frac{\{\delta_3^2 + 4G_0/(nh_n)\}^{1/2} + \delta_3}{2},$$

under the same probability (A.23). which completes the proof of (4.8). To prove (4.9), we use the similar arguments of (A.16) to obtain,

$$\|\hat{m}_p(\mathbf{X}) - m(\mathbf{X})\|_{L_2}^2 \leq \frac{C\{\delta_3^2 + (4KG_0)/n\}}{K},$$

with the probability at least

$$1 - C_7 \exp\left\{ - C_8\left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK\right)\right\}.$$

□

A.5 Lemmas

Lemma 1. *Suppose* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$ *has such a CP decomposition,*

$$\mathbf{A} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \ldots \circ \boldsymbol{\beta}_{r,D} \circ \boldsymbol{\alpha}_r,$$

where $\boldsymbol{\alpha}_r = (\alpha_{r,1}, \cdots, \alpha_{r,K})^\intercal \in \mathbb{R}^K$ *and* $\boldsymbol{\beta}_{r,d} \in \mathbb{R}^{p_d}$ *for* $d = 1, \ldots, D$ *and* $r = 1, \ldots, R$*. If* $\mathbf{u} \in \{(u_1, \cdots, u_K)^\intercal : \sum_{k=1}^{K} \alpha_{r,k} u_k = 0, r = 1, \ldots, R\}$*, then*

$$\sum_{k=1}^{K} A_{\boldsymbol{j},k} u_k = 0, \quad \textit{for} \quad \boldsymbol{j} \in \mathcal{J},$$

where

$$\mathcal{J} = \{(i_1, \cdots, i_D), i_1 = 1, \ldots, p_1, \ldots, \ldots, i_D = 1, \ldots, p_D\}.$$

Proof. This proof is straightforward. For simplicity, for $r = 1, \ldots, R$, let

$$\mathbf{B}_r = \boldsymbol{\beta}_{r,1} \circ \ldots \circ \boldsymbol{\beta}_{r,D}.$$

Since

$$\sum_{k=1}^{K} \alpha_{r,k} u_k = 0,$$

we have

$$\sum_{k=1}^{K} B_{r,\boldsymbol{j}} \alpha_{r,k} u_k = 0, \quad \boldsymbol{j} \in \mathcal{J},$$

where $B_{r,\boldsymbol{j}}$ is $\boldsymbol{j}$-th entry of $\mathbf{B}_r$, $r = 1, \ldots, R$. Therefore,

$$\sum_{k=1}^{K} A_{\boldsymbol{j}.k} u_k = \sum_{k=1}^{K} \sum_{r=1}^{R} B_{r,\boldsymbol{j}} \alpha_{r,k} u_k = \sum_{r=1}^{R} \sum_{k=1}^{K} B_{r,\boldsymbol{j}} \alpha_{r,k} u_k = 0, \quad \boldsymbol{j} \in \mathcal{J}.$$

□

Lemma 2. *Suppose* $\mathbf{U} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ *is a random tensor with its entry* $U_{\boldsymbol{j}} \overset{i.i.d.}{\sim} U(0,1)$*, for* $\boldsymbol{j} \in \mathcal{J}$

and $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$. *Let* $(\Phi(\mathbf{X}))_{\boldsymbol{j},k} = b_k(X_{\boldsymbol{j}})$, *where* $\{b_k(x)\}_{k=1}^K$ *be a B-spline basis,* $x \in [0,1]$. *Under Assumption 1 and 4, if* $\sum_{k=1}^K A_{\boldsymbol{j},k} u_k = 0$ *for* $\boldsymbol{j} \in \mathcal{J}_1 := \mathcal{J}/\{(1,\ldots,1)\}$, *where* $u_k = \int_0^1 b_k(x)\mathrm{d}x$, *then we have*

$i.$

$$C_1 C_\zeta h_n \|\mathbf{A}\|_{HS}^2 \leq \mathbb{E}\{\langle \mathbf{A}, \Phi(\mathbf{X})\rangle^2\} \leq C_2 h_n \|\mathbf{A}\|_{HS}^2, \tag{A.24}$$

and

$ii.$

$$\|\langle \mathbf{A}, \Phi(\mathbf{X})\rangle\|_{\psi_2}^2 \leq C_3 \tilde{h}_n \|\mathbf{A}\|_{HS}^2, \tag{A.25}$$

where C_1, C_2, C_3, C_ζ *are positive constants,* C_ζ *depends on the order of B-spline* ζ, *and*

$$\tilde{h}_n = \max\left\{ \frac{h_n^{1/(-\log h_n)}}{(-2\log h_n)}, h_n \right\}. \tag{A.26}$$

Proof. We will prove the population bound (A.24) at first. Let $\mathbf{A}_{\boldsymbol{j}} = (A_{\boldsymbol{j},1}, \cdots, A_{\boldsymbol{j},K})^\intercal$ for $\boldsymbol{j} \in \mathcal{J}$. By the property of B-spline (see, e.g., De Boor, 1973, 1976) and Assumption 4, for $1 \leq q \leq +\infty$,

$$C_\zeta \|\mathbf{A}_{\boldsymbol{j}}\|_q \leq h_n^{-\frac{1}{q}} \left\| \sum_{k=1}^K \mathbf{A}_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_q \leq C \|\mathbf{A}_{\boldsymbol{j}}\|_q, \tag{A.27}$$

where C_ζ and C are two positive constants and C_ζ depends on the order of B-spline ζ. By the independence and the mean zero restriction for $\boldsymbol{j} \in \mathcal{J}_1$, we have

$$\mathbb{E}[\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2] = \sum_{\boldsymbol{j} \in \mathcal{J}} \mathbb{E}\left[\left\{ \sum_{k=1}^K A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\}^2 \right].$$

Taking $q = 2$ in $(A.27)$ yields

$$C_\zeta h_n \|A_{\boldsymbol{j}}\|_2^2 \leq \mathbb{E}\left[\left\{ \sum_{k=1}^K A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\}^2 \right] \leq C h_n \|\mathbf{A}_{\boldsymbol{j}}\|_2^2,$$

then

$$C_\varsigma h_n \|\mathbf{A}\|_{HS}^2 \leq \mathbb{E}[\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2] \leq C h_n \|\mathbf{A}\|_{HS}^2. \tag{A.28}$$

By Assumption 1, we have

$$C_1 \mathbb{E}[\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2] \leq \mathbb{E}[\langle \mathbf{A}, \Phi(\mathbf{X})\rangle^2] \leq C_4 \mathbb{E}[\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2]. \tag{A.29}$$

It follows from (A.28) and (A.29) that

$$C_1 C_\varsigma h_n \|\mathbf{A}\|_{HS}^2 \leq \mathbb{E}[\langle \mathbf{A}, \Phi(\mathbf{X})\rangle^2] \leq C_2 h_n \|\mathbf{A}\|_{HS}^2,$$

which completes the proof of (A.24).

Now, we will prove the sub-Gaussian norm bound (A.25). Note that

$$\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2} \leq \left\| \sum_{\boldsymbol{j} \in \mathcal{J}_1} \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_{\psi_2} + \left\| \sum_{k=1}^{K} A_{1,\cdots,1,k} b_k(U_{1,\cdots,1}) \right\|_{\psi_2},$$

then

$$\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \leq 2 \left\| \sum_{\boldsymbol{j} \in \mathcal{J}_1} \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_{\psi_2}^2 + 2 \left\| \sum_{k=1}^{K} A_{1,\cdots,1,k} b_k(U_{1,\cdots,1}) \right\|_{\psi_2}^2. \tag{A.30}$$

Using the independence property of $\mathbf{U}$, mean zero restriction of $\mathbf{A}$ and Proposition 2.6.1 of Vershynin (2018), we obtain

$$\left\| \sum_{\boldsymbol{j} \in \mathcal{J}_1} \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_{\psi_2}^2 \leq C_5 \sum_{\boldsymbol{j} \in \mathcal{J}_1} \left\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_{\psi_2}^2. \tag{A.31}$$

It follows from (A.30) and (A.31) that

$$\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \leq 2C_5 \sum_{\boldsymbol{j} \in \mathcal{J}_1} \left\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_{\psi_2}^2 + 2 \left\| \sum_{k=1}^{K} A_{1,\cdots,1,k} b_k(U_{1,\cdots,1}) \right\|_{\psi_2}^2.$$

Therefore,

$$\begin{aligned}
&\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \\
&\quad \le (2C_5+2)\sum_{\boldsymbol{j}\in\mathcal{J}}\Big\|\sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2 + (2C_5+2)\Big\|\sum_{k=1}^{K} A_{1,\cdots,1,k} b_k(U_{1,\cdots,1})\Big\|_{\psi_2}^2 \qquad \text{(A.32)}\\
&\quad = (2C_5+2)\sum_{\boldsymbol{j}\in\mathcal{J}}\Big\|\sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2 .
\end{aligned}$$

We then consider the sub-Gaussian norm of $A_{\boldsymbol{j},k}b_k(U_{\boldsymbol{j}})$. When $q=1$, by (A.27), we have

$$\frac{\|A_{\boldsymbol{j},k}b_k(U_{\boldsymbol{j}})\|_1}{\sqrt{1}} \le 2\frac{\|A_{\boldsymbol{j},k}b_k(U_{\boldsymbol{j}})\|_2}{\sqrt{2}} \le C\sqrt{h_n}\|\mathbf{A}_{\boldsymbol{j}}\|_2. \qquad \text{(A.33)}$$

Similarly, when $q \ge 2$, we obtain

$$\frac{\|A_{\boldsymbol{j},k}b_k(U_{\boldsymbol{j}})\|_q}{\sqrt{q}} \le C\frac{h_n^{1/q}}{\sqrt{q}}\|\mathbf{A}_{\boldsymbol{j}}\|_q \le C\frac{h_n^{1/q}}{\sqrt{q}}\|\mathbf{A}_{\boldsymbol{j}}\|_2. \qquad \text{(A.34)}$$

Since $f(x) = \frac{h_n^{1/x}}{\sqrt{x}}$ get the maximum at $x = -2\log h_n$, then

$$\frac{h_n^{1/q}}{\sqrt{q}}\|\mathbf{A}_{\boldsymbol{j}}\|_2 \le \frac{h_n^{1/(-2\log h_n)}}{(-2\log h_n)^{1/2}}\|\mathbf{A}_{\boldsymbol{j}}\|_2. \qquad \text{(A.35)}$$

Recalling

$$\tilde{h}_n = \max\left\{\frac{h_n^{1/(-\log h_n)}}{(-2\log h_n)}, h_n\right\},$$

and using (A.32)-(A.35), we get

$$\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \le (2C_5+2)\tilde{h}_n C^2\|\mathbf{A}\|_{HS}^2. \qquad \text{(A.36)}$$

Note that for $q \ge 1$,

$$\frac{1}{\sqrt{q}}\big\{\mathbb{E}[|\langle \mathbf{A}, \Phi(\mathbf{X})\rangle|^q]\big\}^{\frac{1}{q}} \le C\frac{1}{\sqrt{q}}\big\{\mathbb{E}[|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle|^q]\big\}^{\frac{1}{q}} \le C\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2},$$

therefore,

$$\|\langle \mathbf{A}, \Phi(\mathbf{X})\rangle\|_{\psi_2}^2 \leq C_3 \tilde{h}_n \|\mathbf{A}\|_{HS}^2,$$

which completes the proof of (A.25). □

Lemma 3. *Let* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$*, and*

$$\mathcal{P} = \left\{ \frac{\mathrm{vec}(\mathbf{A})}{\|\mathbf{A}\|_{HS}} : \sum_{k=1}^{K} A_{\boldsymbol{j},k} u_k = 0, \textit{ for } \boldsymbol{j} \in \mathcal{J}/\{(1,\ldots,1)\},\ \mathrm{rank}(\mathbf{A}) \leq R \right\}, \tag{A.37}$$

where $u_k = \int_0^1 b_k(x)\mathrm{d}x$*. The Gaussian width satisfying*

$$w(\mathcal{P}) \leq C\left(R^{D+1} + R\sum_{d=1}^{D} p_d + RK \right)^{1/2}. \tag{A.38}$$

Proof. By the covering number argument in Lemma 7, we have

$$N(\epsilon, \mathcal{P}, l_2) \leq \left(C_1/\epsilon\right)^{R^{D+1} + R\sum_{i=1}^{D} p_i + RK},$$

where $C_1 = 3D + 4$ is a constant. Suppose $\mathbf{a} \in \mathcal{P}$, then by the Dudley's integral entropy bound (see, e.g., Theorem 3.1 of Koltchinskii (2011)), we obtain

$$\begin{aligned}
\mathbb{E}\sup_{\mathbf{a}\in\mathcal{P}}(\mathbf{a}^\intercal \mathbf{x}) &\leq C_3 \int_0^2 \left\{ \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right) \log\left(\frac{C_1}{x}\right) \right\}^{1/2} \mathrm{d}x \\
&\leq C\left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right)^{1/2}.
\end{aligned}$$

Thus we complete the proof.

□

Lemma 4. *Let* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$ *and suppose* $\mathcal{P}$ *is defined as* (A.37)*. Under Assumption 1, we have*

i.

$$\sup_{\mathbf{A}\in\mathcal{P}}\left|\frac{1}{n}\frac{1}{\mathbb{E}[|\langle\mathbf{A},\Phi(\mathbf{X})\rangle|^2}\sum_{i=1}^{n}\langle\mathbf{A},\Phi(\mathbf{X}_i)\rangle^2-1\right|\leq C_1\tilde{h}_n h_n^{-1}\frac{w(\mathcal{P})}{\sqrt{n}}$$

with probability at least $1-\exp\{-C_2w^2(\mathcal{P})\}$*, where* $w(\mathcal{P})$ *is the Gaussian width ad* $(\Phi(\mathbf{X}))_{\boldsymbol{j},k}=b_k(X_{\boldsymbol{j}})$ *for* $\boldsymbol{j}\in\mathcal{J}$*,* $k=1,\ldots,K$*. Furthermore, suppose* $\|\mathbf{A}\|_{HS}=1$ *and* $n>C\tilde{h}_n^2h_n^{-2}w^2(\mathcal{P})$ *for some* $C>0$*, then with the same probability, we have*

ii.

$$C_3h_n\leq\inf_{\mathbf{A}\in\mathcal{P}}\frac{1}{n}\left|\sum_{i=1}^{n}\langle\mathbf{A},\Phi(\mathbf{X}_i)\rangle\right|^2\leq\sup_{\mathbf{A}\in\mathcal{P}}\frac{1}{n}\left|\sum_{i=1}^{n}\langle\mathbf{A},\Phi(\mathbf{X}_i)\rangle\right|^2\leq C_4h_n. \tag{A.39}$$

Note that if $t\geq w(\mathcal{P})$*, then* t *can be used to replace the above* $w(\mathcal{P})$*.*

Proof. Based on Lemma 2, the following proof is similar to Theorem 12 of Banerjee et al. (2015). We consider the following class of functions

$$F=\left\{f_A:f_A\{\Phi(\mathbf{X})\}=\frac{1}{\sqrt{\mathbb{E}\{|\langle\mathbf{A},\Phi(\mathbf{X})\rangle|^2\}}}\langle\mathbf{A},\Phi(\mathbf{X})\rangle,\mathrm{vec}(\mathbf{A})\in\mathcal{P}\right\}.$$

It is trivial to see that $F\subset S_{L_2}:=\{f:\mathbb{E}[f^2\{\Phi(\mathbf{X})\}]=1\}$. By definition,

$$\sup_{f_A\in F}\|f_A\|_{\psi_2}=\sup_{\mathbf{A}\in\mathcal{P}}\left\|\frac{1}{\sqrt{\mathbb{E}[|\langle\mathbf{A},\Phi(\mathbf{X})\rangle|^2}}\langle\mathbf{A},\Phi(\mathbf{X})\rangle\right\|_{\psi_2},$$

and by Lemma 2, for every $\mathrm{vec}(\mathbf{A})\in\mathcal{P}$,

$$\left\|\frac{1}{\sqrt{\mathbb{E}[|\langle\mathbf{A},\Phi(\mathbf{X})\rangle|^2}}\langle\mathbf{A},\Phi(\mathbf{X})\rangle\right\|_{\psi_2}\leq\kappa_n,$$

where $\kappa_n=C_5\tilde{h}_n^{1/2}h_n^{-1/2}$. Then we obtain

$$\sup_{f_A\in F}\|f_A\|_{\psi_2}\leq\kappa_n.$$

Thus for the γ_2 functionals, we have

$$\gamma_2(F \cap S_{L_2}, \|.\|_{\psi_2}) \leq \kappa_n \gamma_2(F \cap S_{L_2}, \|.\|_{L_2}) \leq C_6 \kappa_n w(\mathcal{P}),$$

where the last inequality follows from Theorem 2.1.1 of Talagrand (2005). By Theorem 10 of Banerjee et al. (2015), we can choose

$$\theta = C_7 C_6 \kappa_n^2 \frac{w(\mathcal{P})}{\sqrt{n}} \geq C_7 \kappa_n \frac{\gamma_2(F \cap S_{L_2}, \|.\|_{\psi_2})}{\sqrt{n}}.$$

As a result, with probability at least $1 - \exp(-C_8 \theta^2 n / \kappa_n^4) = 1 - \exp\{-C_2 w^2(\mathcal{P})\}$, we have

$$\sup_{\mathbf{A} \in \mathcal{P}} \left| \frac{1}{n} \frac{1}{\mathbb{E}[|\langle \mathbf{A}, \Phi(\mathbf{X}) \rangle|^2} \sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle^2 - 1 \right| \leq C_1 \tilde{h}_n h_n^{-1} \frac{w(\mathcal{P})}{\sqrt{n}},$$

where $C_1 = C_7 C_6 C_5^2$ and $C_2 = C_8 C_7^2 C_6^2$ are two positive constants. Suppose $\sqrt{n} > C \tilde{h}_n h_n^{-1} w(\mathcal{P})$ for some $C > 0$, then by Lemma 2, with probability at least $1 - \exp\{-C_2 w^2(\mathcal{P})\}$, we have

$$C_3 h_n \leq \inf_{\mathbf{A} \in \mathcal{P}} \frac{1}{n} \left| \sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle \right|^2 \leq \sup_{\mathbf{A} \in \mathcal{P}} \frac{1}{n} \left| \sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle \right|^2 \leq C_4 h_n.$$

□

Lemma 5. *Suppose* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$, $\operatorname{rank}(\mathbf{A}) \leq R$ *and* $\sum_{k=1}^{K} A_{\boldsymbol{j},k} u_k = 0$ *for* $\boldsymbol{j} \in \mathcal{J}/\{(1, \ldots, 1)\}$, *where* $u_k = \int_0^1 b_k(x) \mathrm{d}x$. *If* $n > C \tilde{h}_n^2 h_n^{-2} \big(R^{D+1} + R \sum_{i=1}^{D} p_i + RK\big)$ *for some constant* $C > 0$, *we then have*

$$\sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle \epsilon_i \leq C_1 \|\mathbf{A}\|_{HS} \left\{ n h_n \left(R^{D+1} + \sum_{i=1}^{D} R p_i + RK \right) \right\}^{1/2}, \tag{A.40}$$

with probability at least

$$1 - C_2 \exp\left\{ - C_3 \left(R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.$$

Proof. We use the notation $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^\intercal$ introduced in the proof of Theorem 4, then the left hand side of (A.40) can be rewritten as

$$\sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle \epsilon_i = (\mathbf{Z}\mathbf{a})^\intercal \boldsymbol{\epsilon}.$$

Consider

$$\Gamma_1 = \left\{ \frac{\mathbf{Z}\mathbf{a}}{\sqrt{\lambda_{\text{Rmax}}(\mathbf{Z}^\intercal \mathbf{Z})}} : \mathbf{a} \in \mathcal{P} \right\},$$

where $\lambda_{\text{Rmax}}(\mathbf{Z}^\intercal \mathbf{Z}) = \sup_{\mathbf{a} \in \mathcal{P}} \|\mathbf{Z}\mathbf{a}\|_2$ and $\mathcal{P}$ is defined as in Lemma 4. By the covering number argument in Lemma 7,

$$N(\epsilon, \mathcal{P}, l_2) \leq \left(\frac{C_4}{\epsilon}\right)^{R^{D+1} + R\sum_{d=1}^{D} p_d + RK},$$

where $C_4 = 3D + 4$ is a constant. Following from the definition of Γ_1, we have

$$N(\epsilon, \Gamma_1, l_2) \leq N(\epsilon, \mathcal{P}, l_2) \leq \left(\frac{C_4}{\epsilon}\right)^{R^{D+1} + R\sum_{i=1}^{D} p_i + RK}.$$

By Assumption 2, $\mathbb{E}\{\exp(t\boldsymbol{\eta}^\intercal \boldsymbol{\epsilon})\} \leq \exp(Ct^2 \|\boldsymbol{\eta}\|^2) \leq \exp(Ct^2)$ for $\boldsymbol{\eta} \in \Gamma_1$. Using the Dudley's integral entropy bound, we have

$$\begin{aligned}
\mathbb{E} \sup_{\boldsymbol{\eta} \in \Gamma_1} (\boldsymbol{\eta}^\intercal \boldsymbol{\epsilon}) &\leq C \int_0^2 \left\{ \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right) \log\left(C_4/\epsilon\right) \right\}^{1/2} \mathrm{d}\epsilon \\
&\leq C_5 \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right)^{1/2}.
\end{aligned}$$

As a direct result (e.g., Theorem 8.1.6 of Vershynin (2018)), we have

$$\begin{aligned}
\sup_{\boldsymbol{\eta} \in \Gamma_1} (\boldsymbol{\eta}^\intercal \boldsymbol{\epsilon}) &\leq C \left[\int_0^2 \left\{ \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right) \log\left(C_4/\epsilon\right) \right\}^{1/2} \mathrm{d}\epsilon + 2t \right] \\
&\leq C_6 \left\{ \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right)^{1/2} + t \right\},
\end{aligned}$$

with probability as least $1 - 2\exp(-t^2)$, which implies

$$(\mathbf{Za})^\intercal \boldsymbol{\epsilon} \leq C_7 \sqrt{\lambda_{\mathrm{Rmax}}(\mathbf{Z}^\intercal \mathbf{Z})} \left(R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2}, \tag{A.41}$$

with probability as least

$$1 - 2\exp\left\{ - \left(R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.$$

Plugging (A.38) and (A.39) into (A.41), we will complete the proof of (A.40). □

Lemma 6. *Suppose* $\int_0^1 f_r(u)\mathrm{d}u = 0$, $r = 1, \ldots, R$ *and Assumption 3 holds. Then there exist* $\alpha_{0r,k}$, $r = 1, \ldots, R$, *such that*

$$\left\| f_r - \sum_{k=1}^{K} \alpha_{0r,k} b_k \right\|_\infty = \mathcal{O}(K^{-\tau}),$$

where $\sum_{k=1}^{K} \alpha_{0r,k} u_k = 0$ *and* $u_k = \int_0^1 b_k(x)\mathrm{d}x$.

Proof. It is a well-known result that for each r, there exists a spline function f_{1r} which can be represented by $\{b_k(x)\}_{k=1}^K$, such that

$$\|f_r - f_{1r}\|_\infty = \mathcal{O}(K^{-\tau}).$$

Let $f_{2r} = f_{1r} - \int_0^1 f_{1r}(u)\mathrm{d}u$, then we have

$$\|f_r - f_{2r}\|_\infty \leq \|f_r - f_{1r}\|_\infty + \left| \int_0^1 f_{1r}(u)\mathrm{d}u \right|.$$

Since

$$\begin{aligned} \left| \int_0^1 f_{1r}(u)du \right| &= \left| \int_0^1 \{f_{1r}(u) - f(u)\}\mathrm{d}u + \int_0^1 f(u)\mathrm{d}u \right| \\ &\leq \|f_r - f_{1r}\|_\infty \\ &= \mathcal{O}(K^{-\tau}), \end{aligned}$$

it is straightforward to get

$$\|f_r - f_{2r}\|_\infty = \mathcal{O}(K^{-\tau}).$$

The proof is completed by noting that f_{2r} is a spline function with mean zero.

□

Lemma 7. *Let* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$. *To simplify the notations, denote* $p_{D+1} = K$. *Let* $\Gamma_2 = \{\mathbf{a} : \|\mathbf{a}\|_2 \leq 1,\ \mathbf{a} = \mathrm{vec}(\mathbf{A}),\ \mathrm{rank}(\mathbf{A}) \leq R\}$. *Then the covering number of* Γ_2 *satisfying*

$$N(\epsilon, \Gamma_2, l_2) \leq \Big(\frac{3D+4}{\epsilon}\Big)^{R^{D+1} + R\sum_{d=1}^{D+1} p_d}. \tag{A.42}$$

Proof. Since the CP decomposition is a special case of the Tucker decomposition (Kolda and Bader, 2009), $\mathbf{A}$ can be represented as

$$\mathbf{A} = \mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D \times_{D+1} \mathbf{B}_{D+1}, \tag{A.43}$$

where $\mathbf{B}_d \in \mathbb{R}^{p_d \times R}$, $d = 1, \ldots, D+1$ and $\mathbf{I} \in \mathbb{R}^{R \times R \ldots \times R}$ is a diagonal tensor of which all the diagonal entries are 1. Let $r_d = \mathrm{rank}(\mathbf{B}_d)$. Through the QR decomposition, we get $\mathbf{A}_d = \mathbf{Q}_d \mathbf{R}_d$, where $\mathbf{Q}_d^\intercal \mathbf{Q}_d = \mathbf{I}_{r_d}$ and $\mathbf{I}_{r_d} \in \mathbb{R}^{r_d \times r_d}$ is the identity matrix. Using the argument in $(A.43)$, we have

$$\begin{aligned}
\mathbf{A} &= (\mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D) \times_{D+1} (\mathbf{Q}_{D+1}\mathbf{R}_{D+1}) \\
&= (\mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D \times_{D+1} \mathbf{R}_{D+1}) \times_{D+1} \mathbf{Q}_{D+1} \\
&= \{(\mathbf{I} \times_{D+1} \mathbf{R}_{D+1}) \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D\} \times_{D+1} \mathbf{Q}_{D+1} \\
&= \{(\mathbf{I} \times_{D+1} \mathbf{R}_{D+1}) \times_1 \mathbf{B}_1 \times_2 \cdots \times_D (\mathbf{Q}_D\mathbf{R}_D)\} \times_{D+1} \mathbf{Q}_{D+1} \\
&= \{(\mathbf{I} \times_D \mathbf{R}_D \times_{D+1} \mathbf{R}_{D+1}) \times_1 \mathbf{B}_1 \times_2 \cdots \times_{D-1} \mathbf{B}_{D-1}\} \times_D \mathbf{Q}_D \times_{D+1} \mathbf{Q}_{D+1} \\
&= \cdots \\
&= (\mathbf{I} \times_1 \mathbf{R}_1 \times_2 \cdots \times_{D+1} \mathbf{R}_{D+1}) \times_1 \mathbf{Q}_1 \times_2 \cdots \times_{D+1} \mathbf{Q}_{D+1}.
\end{aligned}$$

In other words, the CP decomposition will lead a higher-order singular value decomposition (HOSVD)(see,

e.g., De Lathauwer et al., 2000). By Lemma 2 of Rauhut et al. (2017), we obtain

$$N(\epsilon, \Gamma_2, l_2) \leq \Big(\frac{3D+4}{\epsilon}\Big)^{\Pi_{d=1}^{D+1} r_d + \sum_{d=1}^{D+1} p_d r_d}.$$

Therefore (A.42) is shown by noting that $r_d \leq R$ for $d = 1, \ldots, D+1$. □