Broadcasted Nonparametric Tensor Regression

Ya Zhou[1,2], Raymond K. W. Wong[2] and Kejun He[1]

[1]*Institute of Statistics and Big Data, Renmin University of China*
[2]*Department of Statistics, Texas A&M University, College Station, USA*

Abstract

We propose a novel broadcasting idea to model the nonlinearity in tensor regression non-parametrically. Unlike existing non-parametric tensor regression models, the resulting model strikes a good balance between flexibility and interpretability. A penalized estimation and corresponding algorithm are proposed. Our theoretical investigation, which allows the dimensions of the tensor covariate to diverge, indicates that the proposed estimation enjoys desirable convergence rate. Numerical experiments are conducted to confirm the theoretical finding and show that the proposed model has advantage over existing linear counterparts.

Keywords: Nonlinear regression; Tensor low rank; Polynomial splines; Region selection; Elastic-net penalization.

1 Introduction

Recent years have witnessed a massive emergence of tensor data 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 (Zhou et al., 2013). 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. In passing, it is also worth mentioning that regression of a tensor response on a vector covariate (e.g., Sun and Li, 2017; Li and Zhang, 2017; Hu et al., 2019) is also a popular research direction.

Commonly seen are three major types of tensor regression, with different forms of response. The first is scale-on-tensor regression, i.e., the response is a scalar (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 (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 we have a vector response (Miranda et al., 2018). The last one is tensor-on-tensor regression with a tensor output (Hoff, 2015; Lock, 2018; Raskutti et al., 2019).

Most of the aforementioned models make a strong assumption that the tensor covariate is able to predict the response through (known transformations of) linear functions. To

date, very few work go beyond this limitation. On the application side, Zhao et al. (2014) and Hou et al. (2015) used Gaussian processes to model potential nonlinear effects of tensor covariates in video surveillance applications and neuroimaging analyses. Their methods are geared for prediction, but lack for interpretability and theoretical justification. Moreover, the performance of this approach heavily relies upon the choice of kernel function, which is not easy to design for efficiently harnessing the tensor structure.

Another class of methods incorporates nonlinearity through a more explicit function space by imposing low-rank structures on covariates. Kanagawa et al. (2016) considered a regression model with respect to a rank-one tensor covariate, i.e., $\mathbf{X} = \mathbf{x}_1 \circ \mathbf{x}_2 \circ \cdots \circ \mathbf{x}_D$, where $\circ$ denotes the outer product and $\mathbf{x}_d$ is a p_d-dimensional vector, $d = 1, \ldots, D$. Imaizumi and Hayashi (2016) extended this work to a higher-rank tensor and proposed the model

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

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

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

where the Euclidean norm $\|\mathbf{x}_{q,d}\|_2 = 1$ and $\lambda_Q \geq \lambda_{Q-1} \geq \cdots \geq \lambda_1 \geq 0$. When $Q = 1$, (1) recovers the model of Kanagawa et al. (2016). In order to significantly reduce the number of unknown functions to be estimated, a small value of Q is usually recommended. However, in most cases, the tensor covariate is not exactly low-rank, and the rank of the covariate varies from observation to observation within the data set. Furthermore, although the additive form of (1) has reduced the model complexity, simultaneously estimating DR unknown multivariate functions, $g_{d,r}$'s, remains a challenging problem. For example, given a $64 \times 64 \times 64$ 3D-image covariate ($p_1 = p_2 = p_3 = 64$), we need to estimate $3R$ unknown 64-dimensional functions, which will lead to the curse of dimensionality. This aligns with a finding, from Imaizumi and Hayashi (2016), that the asymptotic convergence rate of this model grows exponentially with $\max_d p_d$. Finally, this model is difficult to interpret since the nonlinear modeling is directly built upon the CP representation of the covariates, which may not be unique (Stegeman and Sidiropoulos, 2007).

Therefore, although these existing nonlinear models demonstrate successes in certain applications, they suffer from the curse of dimensionality and possess weak interpretability. In this article, 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 tensor linear model through a predetermined link function g

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

where $\nu \in \mathbb{R}$, $\boldsymbol{\gamma} \in \mathbb{R}^{p_0}$, $\mathbf{B} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$ are unknown parameters, and $\langle \cdot, \cdot \rangle$ denotes the componentwise inner product, i.e., $\langle \mathbf{B}, \mathbf{X} \rangle = \sum_{i_1, \ldots, i_D} B_{i_1, \ldots, i_D} X_{i_1, \ldots, i_D}$. 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 CP 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 identify the regions (entries) of $\mathbf{X}$ that are relevant to predict the response variable.

In many real-world applications, entries within some regions of the tensor (especially images) share similar effects due to certain spatial structures. For examples, Zhou et al. (2013) and Miranda et al. (2018) both provided evidences that brains demonstrate spatially clustered effects on certain diseases. Motivated by these observations and the possibility of nonlinear effects, we propose to "broadcast" similar nonlinear relationships to different entries of the tensor covariate. On a high-level, we model the nonlinear effects by uni-dimensional nonparametric functions, which are supposed to be applied to an individual entry. These uni-dimensional functions are then shared by every entry to indicate the clustered effect. We call this operation of distributing a uni-dimensional function to all entries "broadcasting". Additional scaling coefficients are used to linearly scale the effects of the uni-dimensional functions. Through regularizing these scaling coefficients, we are able to 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 produce more successful results. Therefore we also restrict the scaling coefficient to be low-rank. Combined with multiple broadcasting, the proposed model can produce reasonably complex and interpretable structures, as shown in Section 2.2.

Within the proposed model, all aforementioned ideas are integrated into a (penalized) least squares framework. We develop an alternative updating algorithm as well as the asymptotic rates of convergence for the proposed estimations. Our theory includes tensor linear model (Zhou et al., 2013) as a special case. Unlike Zhou et al. (2013), ours is of high-dimensional nature, which allows $p_1, \ldots, p_D$ to diverge with the sample size. We believe this asymptotic framework is more relevant to many applications, where the data, such as images, involve large values of p_j's as compared to the sample size. To construct the asymptotic analysis, we have provided a novel restricted eigenvalue result. Through a real data example, we demonstrate the power of the proposed broadcasted nonparametric tensor regression. Overall, the proposed method timely responds to a number of growing needs of modeling nonlinearity with interpretable models and rigorous theoretical developments for tensor data.

The rest of this paper is organized as follows. Section 2 introduces the broadcasted nonparametric model. The proposed estimation method with the computational algorithm and the corresponding theoretical results are respectively presented in Sections 3 and 4. The practical performance of the proposed method illustrated via both a simulation study and a real data application can be found in Section 5. The main contributions of this paper are summarized in Section 6 with some concluding remarks. Technical details are provided in a separate online supplemental document.

2 Model

Consider a tensor covariate $\mathbf{X} \in \mathcal{X} := \{(A_{i_1,\ldots,i_D})_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D} : A_{i_1,\ldots,i_D} \in \mathcal{I}\}$, where $\mathcal{I}$ is a compact subset of $\mathbb{R}$. Unless otherwise specified, we assume $\mathcal{I} = [0,1]$ without loss of generality. Throughout this paper, we focus on the general model

$$y = m(\mathbf{X}) + \epsilon, \tag{2}$$

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

2.1 Common nonparametric strategies: curse of dimensionality

As discussed in Section 1, existing work of nonparametric tensor regression is not only lacking in interpretability but also suffering from a slow rate of convergence due to the curse of dimensionality. This issue of dimensionality also occurs if one adopts other common nonparametric regression models, such as additive models, by directly flattening the tensor covariate into a vector. To relate with the standard nonparametric regression methods, we briefly discuss them here, and highlight the issue of dimensionality, which motivates the proposed model in Section 2.2.

One of the most general models for the regression function $m(\cdot)$ is an unstructured (smooth) mapping from $\mathcal{X}$ to $\mathbb{R}$. Despite its flexibility, this model suffers heavily from high 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 an additive form of the regression function (Stone, 1985; Hastie and Tibshirani, 1990; Wood, 2017), i.e.,

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

where $m_{i_1 i_2 \cdots i_D}$'s are unknown uni-dimensional functions. This model however needs to simultaneously estimate $s = \prod_{d=1}^D p_d$ uni-dimensional functions, in which consistent estimation is generally impossible for $s \geq n$. In this case, the sparsity assumption (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 help obtain consistent estimation of the regression function. Nevertheless, general sparse estimators, when applied to a vectorized tensor covariate, would ignore the potential tensor structure and might result in large bias, especially when the sample size n is much smaller than s.

Another common modeling is the single index model (Ichimura, 1993; Horowitz and Härdle, 1996), in the sense that

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

where g is an unknown uni-dimensional function and $a_{i_1,i_2,\ldots,i_D}$'s are s unknown weight parameters. Although there is only one uni-dimensional function to be estimated, this model

involves abundant number of weight parameters, sometimes larger than the sample size. One could also impose sparsity assumption on the weight parameters. The readers are referred to Alquier and Biau (2013), Radchenko (2015), and references therein for more details on this approach. However, similar issue of ignoring tensor structures would also show up. Such problem would be aggravated in more complicated index models such as the additive index model and multiple indices model.

We propose a novel and economical model which makes use of the tensor structure. Our model is closely related to the additive models, but has overcome the aforementioned problems.

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:

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

where f is a uni-dimensional function residing in a function class $\mathcal{H}$ to be specified later, and the scaling s^{-1} is introduced to match with our proposed model (3). In other words, we *broadcast*[1] the same function f to every entry. We formally define the broadcasting operator $\mathcal{B} : \mathcal{H} \times \mathcal{X} \to \mathbb{R}^{p_1 \times \cdots \times p_D}$ by

$$(\mathcal{B}(f, \mathbf{X}))_{i_1 i_2 \ldots i_D} = f(X_{i_1 i_2 \ldots i_D}), \quad \text{for all } i_1, \ldots, i_D.$$

Figure 1 depicts an example of the broadcasting operation. 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 subregions have similar linkages with attention deficit hyperactivity disorder. Miranda et al. (2018) demonstrated that voxels within several subregions 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 the same nonlinear effect on the response is very restrictive. Specifically, in many image data, there are usually only one or a few clusters of entries that are related to the response. Therefore, we move beyond a simple broadcasting structure to achieve more adaptive modeling.

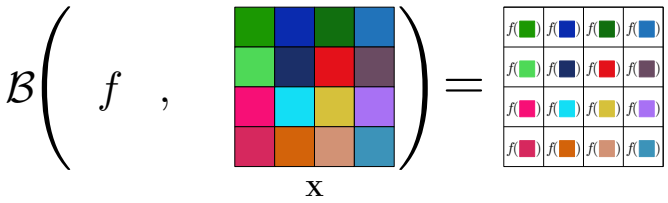

Figure 1: An example of broadcasting operation for a tensor covariate of order 2, i.e., $D = 2$. Different colors represent different possible values that the tensor entries may take.

For any two tensors $\mathbf{A} = (A_{i_1,\ldots,i_D})$ and $\mathbf{B} = (B_{i_1,\ldots,i_D})$ 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 use the (low-rank)

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

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

$$m(\mathbf{X}) = \nu + \frac{1}{s}\sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle \,, \tag{3}$$

where $\nu \in \mathbb{R}$, $\boldsymbol{\beta}_{r,d} \in \mathbb{R}^{p_d}$, and $F_r(\mathbf{X}) = \mathcal{B}(f_r, \mathbf{X})$. Here $f_r \in \mathcal{H}$ admits a nonparametric modeling specified by the (infinite-dimensional) function class $\mathcal{H}$. Following the convention (e.g., Stone, 1985), $\mathcal{H}$ is assumed to be a class of smooth functions 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-one scaling (coefficient) tensor $\boldsymbol{\beta}_{r,1} \circ \cdots \circ \boldsymbol{\beta}_{r,D}$ to linearly scale the effect across different entries. The model is economical since these broadcasted functions are uni-dimensional and these scaling tensors are of rank 1.

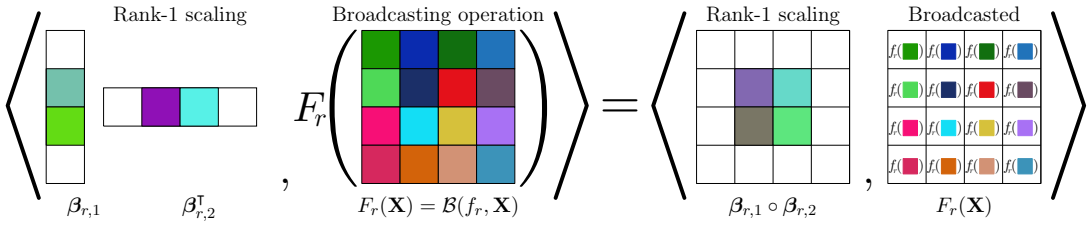

Figure 2: An example of the r-th component in the broadcasted model (3) for $D = 2$, with sparsity in scaling tensor. The white elements in $\boldsymbol{\beta}_{r,1}$, $\boldsymbol{\beta}_{r,2}$, and $\boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2}$ represent zero entries.

If an appropriate sparse estimation is imposed on the scaling tensors, a component can be made specifically concentrated on a subregion of the tensor. We demonstrate the scaling effect in Figure 2. Several components can be combined to characterize different nonlinear effects adapted to different subregions. We give two simple examples of $D = 2$ depicted in Figure 3, where the shaded regions correspond to nonzero entries in the corresponding scaling tensors. In the left panel, there are two rank-one regions (shaded) with different nonlinear functions; in the right panel, there is a rank-two region formed by two scaling tensors with a shared nonlinear effect ($f_1 = f_2$).

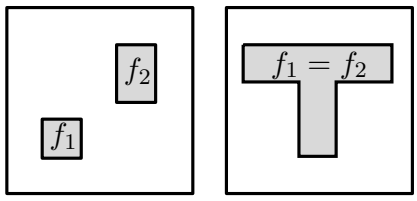

Figure 3: Examples of the broadcasted model (3) for $D = 2$.

Similar to the tensor linear model (Zhou et al., 2013), the parameterization in the proposed model is unidentifiable, i.e., the broadcasted functions and scaling tensors are not uniquely determined. For instance, one can multiply $\boldsymbol{\beta}_{r,1}$ by 10, and divide $\boldsymbol{\beta}_{r,2}$ by 10, while still obtain the same $m(\cdot)$. Another example is a permutation of the components. However, 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. In particular, we are able to directly

study the asymptotic behaviors of the estimations of $m(\cdot)$ in Section 4. For computation, on the other hand, some of these identifiability issues lead to algorithmic instability and so several restrictions are introduced in Section 3 to obtain an efficient algorithm. For a complete discussion on parameter identification, we refer interested readers to Section C of the supplementary material, where sufficient conditions similar to Kruskal's uniqueness condition (Kruskal, 1989) are provided.

3 The proposed estimator and its computation

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.,

$$f_r(x) \approx \sum_{k=1}^{K} \alpha_{r,k} b_k(x), \tag{4}$$

where $\mathbf{b}(x) = (b_1(x), \cdots, b_K(x))^\intercal$ is a vector of B-spline basis functions 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 (3) can be modeled by

$$m(\mathbf{X}) = \nu + \frac{1}{s} \sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \boldsymbol{\alpha}_r, \Phi(\mathbf{X}) \rangle, \tag{5}$$

where $\Phi : \mathcal{X} \to \mathbb{R}^{p_1 \times \cdots \times p_D \times K}$ is defined by $(\Phi(\mathbf{X}))_{i_1,\ldots,i_D,k} = b_k(X_{i_1 \cdots i_D})$. To separate out the constant effect from f_r's, we impose the conditions $\int_0^1 f_r(x)\mathrm{d}x = 0$, $r = 1, \ldots, R$, which lead to

$$\int_0^1 \sum_{k=1}^{K} \alpha_{r,k} b_k(x)\mathrm{d}x = 0, \quad r = 1, \ldots, R. \tag{6}$$

Let $u_k = \int_0^1 b_k(x)\mathrm{d}x$. We consider the following optimization problem

$$\begin{aligned} &\arg\min_{\nu, \mathbf{A}} \sum_{i=1}^{n} \left(y_i - \nu - \frac{1}{s} \langle \mathbf{A}, \Phi(\mathbf{X}_i) \rangle \right)^2 \\ &\text{s.t.} \quad \mathbf{A} = \sum_{r=1}^{R} \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \boldsymbol{\alpha}_r \\ &\qquad \sum_{k=1}^{K} \alpha_{r,k} u_k = 0, \quad r = 1, \ldots, R, \end{aligned} \tag{7}$$

and the estimated regression function as

$$\hat{m}(\mathbf{X}) = \hat{\nu} + \frac{1}{s} \left\langle \hat{\mathbf{A}}, \Phi(\mathbf{X}) \right\rangle,$$

where $(\hat{\mathbf{A}}, \hat{\nu})$ is a solution of (7).

Directly solving (7) is not computationally efficient since it involves too many linear constraints. To further simplify the optimization problem, we remove the constraints by using an equivalent truncated power basis (Ruppert et al., 2003). We let $\{\tilde{b}_k(x)\}_{k=1}^K$ denote the truncated power basis:

$$\tilde{b}_1(x) = 1, \quad \tilde{b}_2(x) = x, \ldots, \tilde{b}_\zeta(x) = x^{\zeta-1},$$
$$\tilde{b}_{\zeta+1}(x) = (x-\xi_2)_+^{\zeta-1}, \ldots, \tilde{b}_K(x) = (x-\xi_{K-\zeta+1})_+^{\zeta-1},$$

where ζ and $(\xi_2, \ldots, \xi_{K-\zeta+1})$ are the order and the interior knots of the aforementioned B-spline. Using these basis functions, we consider the optimization

$$\begin{aligned} &\arg\min_{\tilde{\nu}, \tilde{\mathbf{A}}} \sum_{i=1}^n \left(y_i - \tilde{\nu} - \frac{1}{s} \langle \tilde{\mathbf{A}}, \tilde{\Phi}(\mathbf{X}_i) \rangle \right)^2 \\ &\text{s.t.} \quad \tilde{\mathbf{A}} = \sum_{r=1}^R \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \tilde{\boldsymbol{\alpha}}_r, \end{aligned} \tag{8}$$

where $\tilde{\Phi} : \mathcal{X} \to \mathbb{R}^{p_1 \times \ldots \times p_D \times (K-1)}$ is defined by $(\tilde{\Phi}(\mathbf{X}))_{i_1,\ldots,i_D,k} = \tilde{b}_{k+1}(X_{i_1 \cdots i_D})$, $k = 1, \ldots, K-1$, and $\tilde{\boldsymbol{\alpha}}_r \in \mathbb{R}^{K-1}$ is the vector of coefficients. Compared with (7), the mean zero constraints are removed by reducing one degree of freedom in the basis functions. Theorem 3 in Section B.1 of the supplementary material shows that the optimization (8) results in the same estimated regression function, i.e.,

$$\hat{m}(\mathbf{X}) = \tilde{\nu}_{\text{opt}} + \frac{1}{s} \langle \tilde{\mathbf{A}}_{\text{opt}}, \tilde{\Phi}(\mathbf{X}) \rangle, \tag{9}$$

where $(\tilde{\nu}_{\text{opt}}, \tilde{\mathbf{A}}_{\text{opt}})$ is a solution of (8).

To improve estimation performance (when sample size is relatively small) as well as to enhance interpretability, we add an additional regularization term to the optimization. In particular, a penalized estimation is proposed by solving

$$\begin{aligned} &\arg\min_{\tilde{\nu}, \tilde{\mathbf{A}}} \sum_{i=1}^n \left(y_i - \tilde{\nu} - \frac{1}{s} \langle \tilde{\mathbf{A}}, \tilde{\Phi}(\mathbf{X}_i) \rangle \right)^2 + \sum_{r=1}^R \sum_{d=1}^D \sum_{\ell=1}^{p_d} P_{\boldsymbol{\lambda}}(\beta_{r,d,\ell}) \\ &\text{s.t.} \quad \tilde{\mathbf{A}} = \sum_{r=1}^R \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D} \circ \tilde{\boldsymbol{\alpha}}_r \\ &\qquad \|\tilde{\boldsymbol{\alpha}}_r\|_2^2 = 1, \quad r = 1, \ldots, R, \end{aligned} \tag{10}$$

where $\beta_{r,d,\ell}$ is the ℓ-th element of $\boldsymbol{\beta}_{r,d}$ and $P_{\boldsymbol{\lambda}}(\cdot)$ is the penalty function with tuning parameter $\boldsymbol{\lambda}$. Corresponding estimated regression function $\hat{m}_{\text{pen}}$ can be reconstructed similarly as (9). Typical choices of sparsity-inducing penalty function, within 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 the minimax concave penalty (MCP, Zhang, 2010). In particular, the elastic net can deliver good prediction performance when the predictors are correlated, which usually happens in neuroimaging data (Zhou and Li, 2014). To restrict the nonlinear effects to smaller regions, we consider the elastic-net penalty

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

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ with $\lambda_1 \geq 0$ and $\lambda_2 \in [0, 1]$. Note that the magnitudes of $\boldsymbol{\beta}_{r,1}, \ldots, \boldsymbol{\beta}_{r,D}$ and $\tilde{\boldsymbol{\alpha}}_r$ are not identified. Penalization on $\boldsymbol{\beta}_{r,d}$'s would enlarge $\tilde{\boldsymbol{\alpha}}_r$. In the extreme settings when $\lambda_2 = 0$ (ridge penalty) or 1 (Lasso penalty), the penalization is completely offset by enlarging $\tilde{\boldsymbol{\alpha}}_r$, and becomes redundant. The unit-norm restrictions for $\tilde{\boldsymbol{\alpha}}_r$'s are introduced to prevent these scale issues.

3.2 Computational algorithm

We propose a scale-adjusted block-wise descent algorithm to solve (10). 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 write the squared loss, the penalty, and the whole objective function as

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

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

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 $\mathbf{B}_{-d} = \mathbf{B}_1 \odot \cdots \odot \mathbf{B}_{d-1} \odot \mathbf{B}_{d+1} \odot \cdots \odot \tilde{\mathbf{B}}_{D+1}$, $\mathrm{vec}(\cdot)$ is a vectorization operator, $\odot$ denotes the Khatri–Rao product and $\tilde{\Phi}(\mathbf{X})_{(d)}$ is the mode-d matricization of tensor $\tilde{\Phi}(\mathbf{X})$ (e.g., Kolda and Bader, 2009). 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 $\tilde{\mathbf{B}}_{D+1}$, it can be relaxed to a standard quadratically constrained quadratic program. Therefore, the dual ascent method and second-order cone programming can be used for this block-wise updating.

Of special attention is the magnitude shift among $\boldsymbol{\beta}_{r,d}$'s for $d = 1, \cdots, D$. As an example, we can multiply $\boldsymbol{\beta}_{r,d_1}$ by 10 and divide $\boldsymbol{\beta}_{r,d_2}$ by 10, $d_1 \neq d_2$, without changing the value of the squared loss. This manipulation can, however, change the value of penalty term $G(\boldsymbol{\theta})$. To improve the algorithmic convergence, we propose a rescaling strategy for the elastic-net penalty. For $r = 1, \ldots, R$, we solve the following optimization problem

$$\begin{aligned} \underset{\rho_{r,1}, \ldots, \rho_{r,D}}{\arg\min} & \sum_{d=1}^{D} \left\{ \frac{1}{2}(1 - \lambda_2) \|\rho_{r,d} \boldsymbol{\beta}_{r,d}\|_2^2 + \lambda_2 \|\rho_{r,d} \boldsymbol{\beta}_{r,d}\|_1 \right\} \\ \text{s.t.} \quad & \prod_{d=1}^{D} \rho_{r,d} = 1 \quad \text{and} \quad \rho_{r,d} > 0 \\ & \rho_{r,d} = 1 \text{ if } \boldsymbol{\beta}_{r,d} = \mathbf{0}, \end{aligned} \tag{11}$$

where $\|\cdot\|_1$ is the ℓ_1-norm of a vector, and we use $\hat{\rho}_{r,d}\boldsymbol{\beta}_{r,d}$ to replace $\boldsymbol{\beta}_{r,d}$ at the end of each iterative step of solving (10) (see Algorithm 1), where $\{\hat{\rho}_{r,d} : r = 1, \ldots, R, d = 1, \ldots, D\}$ is the minimizer of (11). This replacement step never increases the objective value (as shown in Proposition 1 below). In particular, as described in Section B.2 of the supplementary material, (11) can be written as a convex problem in an equivalent parametrization. For $\lambda_2 \in (0, 1)$, the method of Lagrange multipliers and Newton's method can be used to solve (11). While for the special boundary cases, i.e., $\lambda_2 \in \{0, 1\}$, we are able to obtain 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 (modulo the sign) of*

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

up to scaling, i.e.,

$$\Theta(\boldsymbol{\theta}) = \{\boldsymbol{\theta}^{\boldsymbol{\rho}} : \boldsymbol{\theta}^{\boldsymbol{\rho}} = (\tilde{\nu}, \mathbf{B}_1\boldsymbol{\rho}_1, \ldots, \mathbf{B}_D\boldsymbol{\rho}_D, \tilde{\mathbf{B}}_{D+1}), \\ \boldsymbol{\rho}_d = \mathrm{diag}(\rho_{1,d}, \ldots, \rho_{R,d}), \prod_{d=1}^{D} \rho_{r,d} = 1, \rho_{r,d} > 0\},$$

where $\mathrm{diag}(\rho_{1,d}, \ldots, \rho_{R,d}) \in \mathbb{R}^{R\times R}$ is a diagonal matrix with diagonal entries $\rho_{1,d}, \ldots, \rho_{R,d}$. We write the solution of (11) as $\hat{\rho}_{r,d}$, $r = 1, \ldots R$, $d = 1, \ldots, D$. Let

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

where $\bar{\mathbf{B}}_d = (\hat{\rho}_{1,d}\boldsymbol{\beta}_{1,d}, \ldots, \hat{\rho}_{R,d}\boldsymbol{\beta}_{R,d})$. 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

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

Proposition 1 shows that $\bar{\boldsymbol{\theta}}$ is the unique minimizer over $\Theta(\boldsymbol{\theta})$. This fixes the scaling indeterminacy, improves the practical convergence property and thus enhances the numerical performance. The numerical comparison between the algorithm with and without the rescaling strategy is presented in Section 5. Besides, the convergence of Algorithm 1 is shown in Proposition 2 and its proof is deferred to Section B.4 of the supplementary material.

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 stationary points of $LG(\boldsymbol{\theta})$ are isolated. Then the sequence $\boldsymbol{\theta}^{(t)}$ generated by Algorithm I converges to a stationary point of $LG(\boldsymbol{\theta})$.*

Due to space constraint, the strategy for initializing the algorithm is presented in Section D of the supplementary material. As for tuning parameter selection, cross-validation can be adopted. However, it is computationally expensive. For simplicity, we adopt the validation method in our numerical experiments.

Algorithm 1: Scale-adjusted block relaxation algorithm.

Input : $\boldsymbol{\theta}^{(0)} = \big(\tilde{\nu}^{(0)}, \mathbf{B}_1^{(0)}, \ldots, \mathbf{B}_D^{(0)}, \tilde{\mathbf{B}}_{D+1}^{(0)}\big)$, $\epsilon > 0$ and $t = 0$.

repeat

  for *d from* $1, \ldots, D$ **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

  $\tilde{\mathbf{B}}_{D+1}^{(t+1)} = \arg\min_{\tilde{\mathbf{B}}_{D+1}} LG(\tilde{\nu}^{(t)}, \mathbf{B}_1^{(t+1)}, \ldots, \mathbf{B}_D^{(t+1)}, \tilde{\mathbf{B}}_{D+1})$;

  $\tilde{\nu}^{(t+1)} = \arg\min_{\tilde{\nu}} LG\big(\tilde{\nu}, \mathbf{B}_1^{(t+1)}, \ldots, \mathbf{B}_D^{(t+1)}, \tilde{\mathbf{B}}_{D+1}^{(t+1)}\big)$;

  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 $\hat{\rho}_{r,d}^{(t+1)}$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, are obtained from (11);

  $t = t + 1$;

until $-LG(\boldsymbol{\theta}^{(t+1)}) + LG(\boldsymbol{\theta}^{(t)}) \leq \epsilon$.

Output: $\hat{\boldsymbol{\theta}}_{\text{pen}} = \boldsymbol{\theta}^{(t)}$.

4 Theoretical study

Throughout the theoretical analysis, we assume that the true regression function $m_0(\mathbf{X})$ 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 $F_{0r} = \mathcal{B}(f_{0r}, \mathbf{X})$ with $\int_0^1 f_{0r}(x)\mathrm{d}x = 0$, $f_{0r} \in \mathcal{H}$, $r = 1, \ldots, R_0$, and $\mathcal{H}$ is the function class specified in Assumption 3.

We use C and C with subscripts to represent generic constants that may change values from line to line. In our analysis, we need the following regularity assumptions.

Assumption 1. *The covariate tensor* $\mathbf{X} \in [0,1]^{p_1 \times \cdots \times p_D}$ *has a continuous probability 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* $S_1, S_2 > 0$ *such that* $S_1 \leq g(\mathbf{x}) \leq S_2$ *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 there exists a positive constant* S *such that*

$$(\mathbb{E}|X|^p)^{1/p} \leq S\sqrt{p}, \quad \textit{for all } p \geq 1.$$

The minimum value of S *is the sub-Gaussian norm of* X*, denoted by* $\|X\|_{\psi_2}$.

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 sub-Gaussian with mean 0 and sub-Gaussian norm* $\sigma < \infty$.

Assumption 3. *The true broadcasted functions $f_{0r} \in \mathcal{H}$, $r = 1, \ldots, R_0$. Here $\mathcal{H}$ is the space of functions from $[0,1]$ to $\mathbb{R}$ satisfying the Hölder condition of order ω, i.e.,*

$$\mathcal{H} = \big\{g : \exists\, C \in (0,\infty) \text{ s.t. } |g^{(\iota)}(x_1) - g^{(\iota)}(x_2)| \le C|x_1 - x_2|^{\omega},\ \forall\ x_1, x_2 \in [0,1]\big\},$$

where ι is a nonnegative integer and $g^{(\iota)}$ is the ι-th derivative of g, such that $\omega \in (0,1]$ and $\tau = \iota + \omega > 1/2$.

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

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

Assumptions 1, 3, and 4 are common in nonparametric regression models. In particular, Assumptions 3 and 4 regularize the space where the true broadcasted functions lie in and guarantee that they can be approximated well 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

$$\Big\| f_{0r} - \sum_{k=1}^{K} \alpha_{0r,k} b_k \Big\|_{\infty} = \mathcal{O}(K^{-\tau}), \tag{12}$$

where the L_∞-norm of a uni-dimensional function f is defined as $\|f\|_\infty = \sup_x |f(x)|$. Although we assume $\int_0^1 f_{0r}(x)\mathrm{d}x = 0$, Lemma A.6 (in the supplementary material) still implies that there are $\boldsymbol{\alpha}_{0,r}$, $r = 1, \ldots, R$, satisfying (12) with

$$\sum_{k=1}^{K} \int_0^1 \alpha_{0r,k} b_k(u)\mathrm{d}u = 0. \tag{13}$$

Despite this mild difference in parameter identification, similar assumptions can be found in Zhou et al. (1998) and Huang et al. (2010). Besides, the sub-Gaussianity condition in Assumption 2 is now a standard tail condition of the error.

We present the convergence rates of $\hat{m}(\mathbf{X})$. The function norm for a tensor function m is defined as $\|m\| = [\{\mathbb{E}_{\mathbf{X}} m^2(\mathbf{X})\}]^{1/2}$, which is equivalent to the L_2-norm, $\|m\|_{L_2} = \{\int m^2(\mathbf{X})\mathrm{d}\mathbf{X}\}^{1/2}$, due to Assumption 1. To simplify the notations, we write $\mathbf{B}_{0r} = \boldsymbol{\beta}_{0r,1} \circ \ldots \circ \boldsymbol{\beta}_{0r,D}$. Theorem 1 shows the convergence rates of the unpenalized estimator based on (7), where the parameters p_i, K, R and R_0 are allowed to go to infinity with the sample size n.

Theorem 1. *Suppose $\hat{m}(\mathbf{X})$ is the estimated regression function reconstructed from (7). If Assumptions 1–4 hold, $R \ge R_0$, and $n > C h_n^{-2-2/(\log h_n)} (\log^{-2} h_n)\big(R^{D+1} + \sum_{i=1}^{D} R p_i + RK\big)$ for some $C > 0$, then we have the following result:*

$$\begin{aligned} \|\hat{m} - m_0\|^2 = \mathcal{O}_{\mathrm{p}}\left(\frac{R^{D+1} + \sum_{i=1}^{D} R p_i + RK}{n}\right) \\ + \mathcal{O}_{\mathrm{p}}\left(\left\{\frac{\sum_{r=1}^{R_0} \|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s}\right\}^2 \frac{1}{K^{2\tau}}\right). \end{aligned} \tag{14}$$

The proof of Theorem 1 is not straightforward. To see this, if we discard the low-rank and broadcasting structure of the proposed model (3), we can rewrite the regression function as a nonparametric additive model, and vectorize the basis tensor and its coefficients in (5), i.e.,

$$m(\mathbf{X}_i) = \nu + \mathbf{z}_i^\intercal \mathbf{a},$$

where $\mathbf{z}_i = \text{vec}\{\Phi(\mathbf{X}_i)\}$ and $\mathbf{a} = \text{vec}(\mathbf{A})$. 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", i.e., $\mathbf{Z}^\intercal\mathbf{Z}/n$, where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\intercal$. For a fixed number of predictors, 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 result in a diminishing lower bound at an exponential rate of s, when the number of predictors s goes to infinity with the sample size n (see, e.g., Chen et al., 2018). Therefore, a new study of eigenvalue bounds is needed to carefully harness the model structure, in particular the low-CP-rank structure of $\mathbf{A}$ (see (7)). In our proof, we obtain well-controlled bounds of the restricted eigenvalues over a set of low-CP-rank tensor, that holds with high probability when the sample size is of order $K^2\big(R^{D+1} + \sum_{i=1}^D Rp_i + RK\big)$; see the more precise version (A.11) in the supplementary material. The resulting eigenvalue bounds fill in the gap, with a reasonable sample size dependence.

Roughly speaking, the first term and the second term in (14) correspond to the estimation error and the approximation error, respectively. We can see that the estimation error roughly scales with the number of effective parameters in the model. Thus, we believe the derived rate is very close to the optimal one, if not the same. 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 best rates of convergence. Let

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

If $\delta_1\delta_2^{-1/(2\tau+1)} R^{-2\tau/(2\tau+1)} \geq n^{1/(2\tau+1)}$, the best rate 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 $K \asymp (n\delta_2/R)^{1/(2\tau+1)}$ results in the best rate $(R/n)^{2\tau/(2\tau+1)}\delta_2^{1/(2\tau+1)}$. One special case is that when p_i, R and R_0 do not grow with n, choosing $K \asymp n^{1/(2\tau+1)}$ leads to 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 preferred, especially when the number of predictors are large. Theorem 2 shows the rates of convergence of the penalized method. 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 $\hat{m}_{\text{pen}}(\mathbf{X})$ is the estimated regression function reconstructed from (10). If Assumptions 1–4 hold, $R \geq R_0$ and $n > Ch_n^{-2-2/(\log h_n)}(\log^{-2} h_n)\big(R^{D+1} + \sum_{i=1}^D Rp_i + RK\big)$*

for some $C > 0$, then

$$\|\hat{m}_{\text{pen}} - m_0\|^2 \leq \frac{C_1\{\delta_3^2 + (4KG_0)/n\}}{K} \tag{15}$$

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 (A.21) and

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

Compared with Theorem 1, Theorem 2 has an additional 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 terms of rates of convergence.

5 Experiments

To evaluate the empirical performance of the proposed broadcasted nonparametric tensor regression (BroadcasTR), we compared BroadcasTR with two alternatives upon both synthetic (Section 5.1) and real data sets (Section 5.2). These alternatives are (i) elastic-net regression on the vectorized tensor predictor (ENetR) (Zou and Hastie, 2005) and (ii) tensor linear regression (TLR) (Zhou et al., 2013). Throughout the numerical experiments, ENetR and TLR were implemented by the R package "glmnet" (Friedman et al., 2010) and the MATLAB toolbox "TensorReg" (Zhou et al., 2013) respectively. Since the proposed rescaling strategy (11) can be applied to the computation of tensor linear regression, we also considered this algorithmic modification in our study. To distinguish this modification, we use TLR and TLR-rescaled to represent the algorithm of Zhou et al. (2013) and our algorithm with scaling strategy respectively.

We aim to evaluate estimation and prediction performance as well as region selection of these methods. To identify entry-wise contribution of the covariate tensor, we note that the estimated regression function m of the above methods can be expressed as an additive form $\hat{\nu} + \sum_{i_1,\ldots,i_D} \hat{m}_{i_1,\ldots,i_D}(X_{i_1,\ldots,i_D})$, where $\int_0^1 \hat{m}_{i_1,\ldots,i_D}(x)\mathrm{d}x = 0$, and so the entry-wise effect can be summarized by the L_2-norm $\|\hat{m}_{i_1,\ldots,i_D}\|_{L_2} = \{\int_0^1 \hat{m}^2_{i_1,\ldots,i_D}(x)\mathrm{d}x\}^{1/2}$. More specifically, $\hat{m}_{i_1,\ldots,i_D}$ is a nonlinear function for BroadcasTR, and a linear function for other alternatives. Putting together these entrywise nonlinear effects, we obtain a tensor of dimension $p_1 \times \cdots \times p_D$ with the $(i_1, \ldots, i_D)$-th element being $\|\hat{m}_{i_1,\ldots,i_D}\|_{L_2}$. In below, it is called the norm tensor of the corresponding tensor regression method. We use this norm tensor to indicate important subregions identified by all methods.

5.1 Synthetic data

Similar to Zhou et al. (2013), we fix the dimension of $\mathbf{X}$ to be 64×64. In our simulation study, we consider four different regression functions:

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

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

$$\text{Case 3: } y = m_3(\mathbf{X}) + \epsilon_3 = 1 + \langle \mathbf{B}_3, F_1(\mathbf{X}) \rangle + \epsilon_3,$$

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

where $F_1, F_2 : [0,1]^{64\times 64} \to \mathbb{R}^{64\times 64}$ satisfy

$$(F_1(\mathbf{X}))_{i_1,i_2} = f_1(X_{i_1,i_2}) = X_{i_1,i_2} + 0.6 \sin\{2\pi(X_{i_1,i_2} - 0.5)^2\},$$
$$(F_2(\mathbf{X}))_{i_1,i_2} = f_2(X_{i_1,i_2}) = X_{i_1,i_2} + 0.3 \cos(2\pi X_{i_1,i_2}),$$

for $i_1 = 1, \ldots, 64$ and $i_2 = 1, \ldots, 64$. The scaling matrices $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$, $\mathbf{B}_{41}$ and $\mathbf{B}_{42}$ are binary and depicted in the first column of Figure 4. These regression functions are used to illustrate four different situations: (1) a linear model with one important rank-two subregion; (2) a nonlinear model with one important rank-four subregion; (3) a nonlinear model with two separated important rank-two subregions that share the same nonlinearity; and (4) a low rank nonlinear model with two separated important rank-two subregions that show different nonlinearities.

For each Case j, the covariate tensor $\mathbf{X}$ and the error ε_j were generated such that $X_{i_1,i_2} \sim$ Uniform$[0,1]$, $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ independently across all i_1, i_2. The parameter σ_j was set to be 10% of the standard deviation of entries of $m_j(\mathbf{X})$. We generated 50 simulated data sets independently for each setting of sample size $n = 500$, 750, 1000. Each simulated data set was then split into two separate subsets: a training set with 80% data and a validation set with 20% data. The tuning parameters of the underlying methods were selected by minimizing the validation error:

$$\frac{1}{n_{\text{valid}}} \sum_{i=1}^{n_{\text{valid}}} (y_{\text{valid},i} - \hat{y}_{\text{valid},i})^2,$$

over grids of corresponding tuning parameters, where n_{valid} is the size of the validation set, $\hat{y}_{\text{valid},i}$ is the prediction value of the i-th observation $y_{\text{valid},i}$ in the validation set. These grids can be found in Section E of the supplementary material.

To evaluate the estimation performance, we define the integrated squared error (ISE)

$$\text{ISE} = \|\hat{m} - m_0\|_{L_2}^2,$$

where m_0 and $\hat{m}$ are the true function and a generic estimated function respectively. We note that $\|\hat{m} - m_0\|_{L_2} = \|\hat{m} - m_0\|$ (which we adopt in Section 4), due to the uniform distribution of $\mathbf{X}$. The average ISEs of the proposed and comparative methods are summarized in Table 1. For the nonlinear situations, i.e., Cases 2–4, it is shown that the proposed BroadcasTR outperforms the other methods significantly. In particular, BroadcasTR reduces the average ISEs by 86%–99% in Case 2, 65%–99% in Case 3, and 74%–98% in Case 4. As for Case 1, which is the linear setting and in favor of the alternative methods, BroadcasTR remains competitive. It performs better than both TLR and ENetR by showing 54%–99% reduction in the average ISEs, and is slightly inferior to TLR-rescaled. Besides, TLR-rescaled performs much better than TLR although they originate from the same penalized regression. This indicates that the proposed rescaling strategy leads to significant improvements. Furthermore, the accuracy

Table 1: Estimation performance in synthetic data. Reported are the averages of ISEs and its standard deviation (in parenthesis) based on 50 data replications. In the first column, n is the total sample size, of which 20% were kept for validation.

n	Case	TLR	TLR-rescaled	ENetR	BroadcasTR
500	1	0.251 (0.054)	**0.066** (0.012)	16.559 (0.722)	0.092 (0.018)
	2	24.422 (2.531)	22.282 (1.683)	31.255 (1.261)	**3.185** (1.843)
	3	76.331 (5.371)	71.915 (6.543)	75.425 (1.770)	**25.228** (5.412)
	4	90.285 (7.081)	89.560 (6.742)	89.253 (3.060)	**23.647** (6.330)
750	1	0.115 (0.024)	**0.038** (0.007)	14.747 (0.606)	0.053 (0.009)
	2	20.899 (1.444)	17.039 (1.665)	30.599 (0.897)	**0.640** (0.186)
	3	71.609 (9.236)	53.298 (4.244)	73.936 (2.681)	**3.373** (2.358)
	4	80.444 (12.828)	57.946 (4.459)	86.782 (3.576)	**4.166** (2.807)
1000	1	0.077 (0.013)	**0.028** (0.007)	10.122 (1.021)	0.034 (0.006)
	2	18.815 (2.133)	15.212 (0.969)	29.718(0.637)	**0.315** (0.046)
	3	65.387 (10.013)	45.331 (2.868)	71.729 (2.130)	**0.962** (0.201)
	4	63.848 (10.967)	51.381 (2.583)	84.549 (2.355)	**1.291** (0.437)

of estimation increases with the sample size for the proposed BroadcasTR, which is consistent with our asymptotic analysis.

The important subregions for BroadcasTR, TLR, TLR-rescaled, and ENetR are identified by their norm tensors (matrices) as defined above. For each method, the norm tensor with the median ISE among 50 simulated datasets of $n = 1000$ was depicted in Figure 4. It shows that BroadcasTR, TLR and TLR-rescaled have similar region selection result for Case 1 (a low-rank model with linear effects), whereas BroadcasTR is much better than TLR and TLR-rescaled for Cases 2–4 (low-rank models with nonlinearity). In all cases, ENetR is unable to identify the import regions, which is an empirical evidence that incorporating the tensor low-rank structure can improve region selection, hence enhancing interpretability. We also present the region identification performance of BroadcasTR for smaller sample sizes ($n = 500$ or 750) in Figure 5. It is not surprising to see that when the sample size increases, the accuracy of identified regions of our proposed method improves.

5.2 Monkey's electrocorticography data

We also evaluated various methods on a publicly available monkey's electrocorticography (ECoG) data set (Shimoda et al., 2012). The corresponding tensor covariate is a preprocessed ECoG signal (Shimoda et al., 2012), organized as a third order tensor of dimensions $64\times10\times10$ (channel $\times$ frequency $\times$ time), and the response variable is the movement distance of a monkey's left shoulder marker along a particular direction. The data preprocessing procedure is similar to Shimoda et al. (2012) which tracks 15-minute experiments. Corresponding details are given in Section F of the supplementary material. Following Hou et al. (2015), we chose 10000 observations of the whole data set starting from the second minute of the experiments. The data set was then randomly split into three different subsets, i.e., a training set, a validation set, and a test set, of size 4000, 1000, and 5000 respectively.

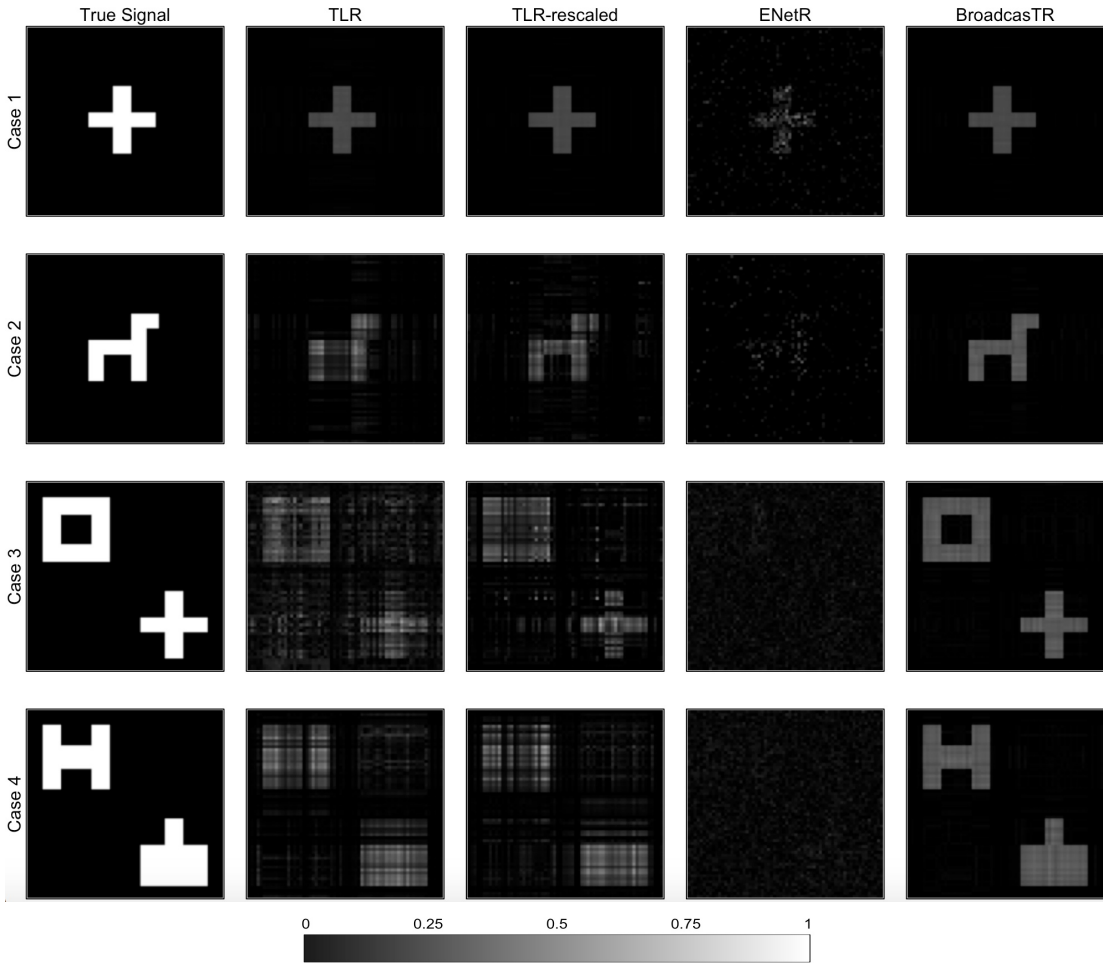

Figure 4: Region selection of TLR, TLR-rescaled, ENetR, and BroadcasTR for $n = 1000$, of which 20% were for validation. The first column presents the true scaling tensors, which are $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3$ and $\mathbf{B}_{41} + \mathbf{B}_{42}$ for Cases 1, 2, 3 and 4, respectively. The rest four columns depict the estimated norm tensor with median ISEs of the comparative and proposed methods. Columns from left to right respectively correspond to TLR, TLR-rescaled, ENetR, and BroadcasTR. The plots in all columns share the same color scheme as shown in the color bar at the bottom.

The grids of tuning parameters we used for the real data can be found in Section E of the supplementary material. To measure the performance, we use mean squared prediction error (MSPE)

$$\text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_{\text{test},i} - \hat{y}_{\text{test},i})^2,$$

where n_{test} is the size of the test set, $\hat{y}_{\text{test},i}$ is the prediction value of the i-th observed value $y_{\text{test},i}$ in the test set. We repeated the fittings for 10 random splittings. The average MSPE over these 10 fittings are reported in Table 2, from which we can see BroadcasTR performs significantly better than the others in prediction.

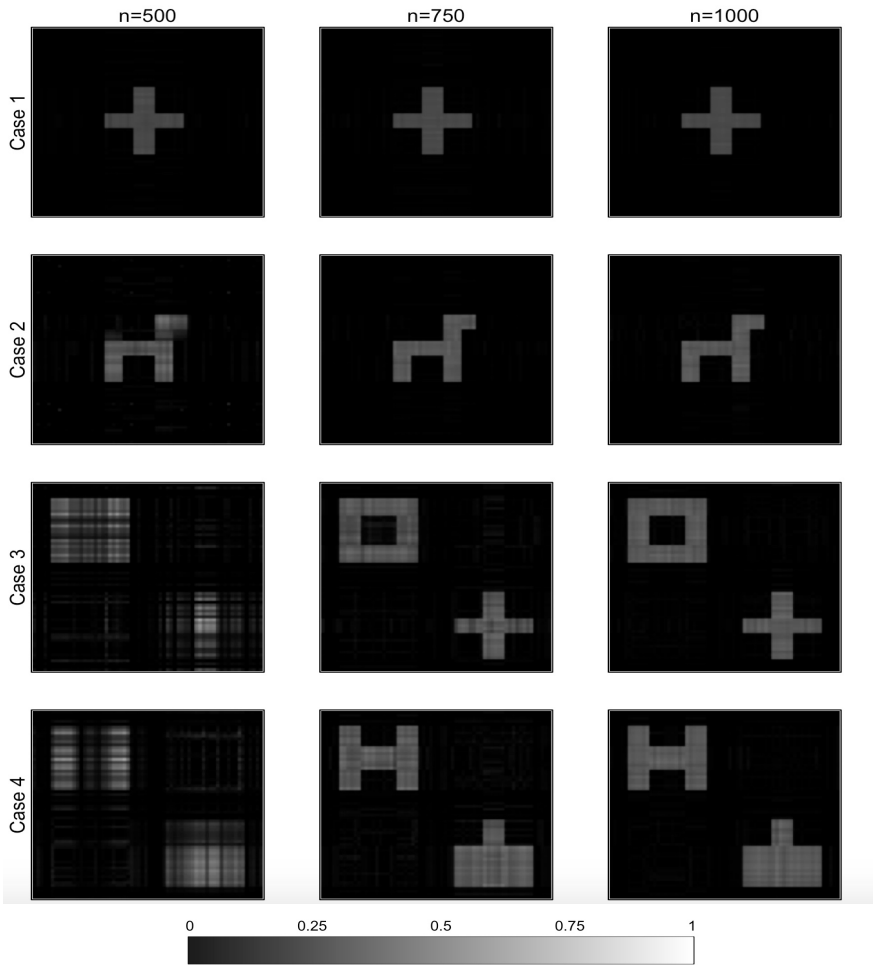

Figure 5: Region selection of BroadcasTR for Cases 1–4, with various sample size $n = 500$, 750, 1000 (where 20% data are used for validation). All plots share the same color scheme as shown in the color bar at the bottom.

6 Conclusion

In this paper, we have proposed a broadcasted model to study the problem of nonlinear regressions with tensor covariates. The curse of dimensionality is tamed by simultaneously utilizing the low-rank tensor structure and broadcasting a uni-dimensional function within each component. With a regularized estimation, the proposed model shows the advantages of improved prediction performance and identifying the important regions on the tensor covariates. Moreover, the convergence rates of the estimator are derived, based on a novel restricted eigenvalue result. We use both synthetic and real data sets to evaluate the empirical performance of the proposed broadcasted nonparametric regression model with some comparison methods, and the results confirm our theoretical findings. The convergence of our proposed estimation algorithm can be guaranteed, with the proposed generalized rescaling strategy. Although our method has concentrated on the problem of continuous and univariate response

Table 2: Prediction performance on the monkey's electrocorticography data. Reported are averages of MPSE and its standard deviation (in parenthesis) based on 10 random splittings.

Data	TLR	TLR-rescaled	ENetR	BroadcasTR
Monkey	3.1703 (0.0418)	3.0923 (0.0699)	3.1256 (0.0431)	**2.5887** (0.0802)

throughout the paper, it is not difficult to generalize it to a classification paradigm or models with multivariate responses. When one dimension of the tensor covariates is ultra-high, using the low-rank structure alone may not be sufficient to obtain a consistent estimation with promising prediction performance. Thus one interesting future research topic is to develop an alternative method incorporating the low-rank tensor with the entry-wise or slice-wise sparsity. Moreover, modeling the regression function by index models with dimension reduction techniques of tensor covariates is also of interest, and needs further investigation.

A Asymptotic study

A.1 Notations

To simplify the notations, we let

$$\mathcal{J} = \{\boldsymbol{j} = (i_1, \cdots, i_D) : 1 \leq i_d \leq p_d,\, d = 1, \ldots, D\}. \tag{A.1}$$

By noting that $s = \Pi_{d=1}^{D} p_d$, we have the cardinality $|\mathcal{J}| = s$. The Hilbert-Schmidt norm of a generic tensor $\mathbf{A}$ is defined as $\|\mathbf{A}\|_{HS} = \langle \mathbf{A}, \mathbf{A} \rangle^{1/2}$.

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.

For convenience, we use a mapping $\Omega : \mathbb{R}^{p_1\times\ldots\times p_D\times K}\times\mathbb{R}\to\mathbb{R}^{p_1\times\ldots\times p_D\times K}$ to represent the operator of absorbing the constant into the coefficients of B-spline basis for the first predictor. More precisely, Ω is defined by

$$\mathbf{A}^\flat = \Omega(\mathbf{A}, \nu), \tag{A.2}$$

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

$$\nu + \frac{1}{s}\langle\mathbf{A}, \Phi(\mathbf{X})\rangle = \frac{1}{s}\langle\mathbf{A}^\flat, \Phi(\mathbf{X})\rangle. \tag{A.3}$$

This property simplifies the development of the asymptotic theory since $\mathbf{A}^\flat$ still enjoys a CP structure.

We also write $\mathbf{A}_0 = \sum_{r=1}^{R_0}\mathbf{B}_{0r}\circ\boldsymbol{\alpha}_{0r}$, $r = 1,\ldots,R$, where $\boldsymbol{\alpha}_{0r}$ satisfies (12) and (13). We define

$$\tilde{h}_n = \max\left\{\frac{h_n^{1/(-\log h_n)}}{(-2\log h_n)}, h_n\right\}. \tag{A.4}$$

With this definition, we have

$$\tilde{h}_n^2 h_n^{-2} \asymp h_n^{-2-2/(\log h_n)}(\log^{-2} h_n), \tag{A.5}$$

which is a quantity presented in the sample size requirements of Theorems 1 and 2.

A.2 Proof of Theorem 1

Suppose $(\hat{\mathbf{A}}, \hat{\nu})$ is a solution of (7), which gives

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

Denoting $\hat{\mathbf{A}}^\flat = \Omega(\hat{\mathbf{A}}, \hat{\nu})$ and $\mathbf{A}_0^\flat = \Omega(\mathbf{A}_0, \nu_0)$ where the operator Ω is defined in (A.2), the aforementioned inequality is equivalent to

$$\sum_{i=1}^n\left(y_i - \frac{1}{s}\langle\hat{\mathbf{A}}^\flat, \Phi(\mathbf{X}_i)\rangle\right)^2 \le \sum_{i=1}^n\left(y_i - \frac{1}{s}\langle\mathbf{A}_0^\flat, \Phi(\mathbf{X}_i)\rangle\right)^2. \tag{A.6}$$

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

$$\mathbf{Z} = (\mathbf{z}_1,\ldots,\mathbf{z}_n)^\intercal \in \mathbb{R}^{n\times sK}, \tag{A.7}$$

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.6) and working out the squares, we obtain

$$\frac{1}{s^2}\|\mathbf{Z}\mathbf{a}^\sharp\|_2^2 \le 2\left\langle\frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp, \boldsymbol{\epsilon}\right\rangle + 2\left\langle\frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp, \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat\right\rangle, \tag{A.8}$$

where $\mathbf{y} = (y_1,\cdots,y_n)^\intercal$. By Lemma A.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 $\text{rank}(\mathbf{A}_0^\flat) \le R_0 + 1$, $\text{rank}(\hat{\mathbf{A}}^\flat) \le R + 1$, it is trivial to see $\text{rank}(\mathbf{A}^\sharp) \le R_0 + R + 2$. 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.8).

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

$$\mathcal{P}_1 = \left\{ \frac{\text{vec}(\mathbf{A})}{\|\mathbf{A}\|_{HS}} : \sum_{k=1}^{K} A_{\boldsymbol{j},k} u_k = 0, \text{ for } \boldsymbol{j} \in \mathcal{J}/\{(1,\ldots,1)\}, \text{rank}(\mathbf{A}) \le R_1 \right\}, \tag{A.9}$$

where

$$R_1 = R + R_0 + 2 \le 2R + 2. \tag{A.10}$$

By Lemma A.4, if $n > C\tilde{h}_n^2 h_n^{-2} w^2(\mathcal{P}_1)$ for some $C > 0$,

$$C_1 n h_n \|\mathbf{a}^\sharp\|_2^2 \le \|\mathbf{Z}\mathbf{a}^\sharp\|_2^2 \le C_2 n h_n \|\mathbf{a}^\sharp\|_2^2, \tag{A.11}$$

with probability as least $1 - 2\exp\{-C_3 w^2(\mathcal{P}_1)\}$. By (A.10) and Lemma A.3, the Gaussian width

$$w(\mathcal{P}_1) \le C\left(R_1^{D+1} + R_1 \sum_{d=1}^{D} p_d + R_1 K \right)^{1/2} \le C_4 \left(R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2}. \tag{A.12}$$

In the following, 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.11) holds with probability tending to 1. By (A.10), (A.11), and Lemma A.5, we have the following upper bound

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

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

$$\begin{aligned}
\left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\|_2^2 &= \sum_{i=1}^{n} \left| \frac{1}{s} \sum_{r=1}^{R_0} \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i) \rangle - \langle \mathbf{A}_0, \Phi(\mathbf{X}_i) \rangle \right|^2 \\
&\le \sum_{i=1}^{n} \left(\frac{1}{s} \sum_{r=1}^{R_0} \left| \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i) \rangle - \langle \mathbf{B}_{0r} \circ \boldsymbol{\alpha}_{0r}, \Phi(\mathbf{X}_i) \rangle \right| \right)^2 \\
&\le \sum_{i=1}^{n} \left\{ \frac{1}{s} \sum_{r=1}^{R_0} \frac{C}{K^\tau} \|\text{vec}(\mathbf{B}_{0r})\|_1 \right\}^2 \\
&= C \left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(\mathbf{B}_{0r})\|_1}{s} \right\}^2 \frac{n}{K^{2\tau}}.
\end{aligned} \tag{A.14}$$

Using the Cauchy-Schwarz inequality and (A.11), it is shown that

$$\begin{aligned}
\left\langle \frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp, \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\rangle &\le \left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\|_2 \left\| \frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp \right\|_2 \\
&= \frac{1}{s}\|\mathbf{Z}\mathbf{a}^\sharp\|_2 \, \mathcal{O}_{\text{p}}\left(\left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{\sqrt{n}}{K^\tau} \right) \\
&= \frac{1}{s}\|\mathbf{a}^\sharp\|_2 \, \mathcal{O}_{\text{p}}\left(\left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{n\sqrt{h_n}}{K^\tau} \right).
\end{aligned} \tag{A.15}$$

Finally, plugging (A.13) and (A.15) into (A.8), we get

$$\begin{aligned}\frac{1}{s^2}\|\mathbf{Z}\mathbf{a}^\sharp\|_2^2 \le \frac{1}{s}\|\mathbf{a}^\sharp\|_2\,\mathcal{O}_{\mathrm{p}}\left(\left\{nh_n\left(R^{D+1}+\sum_{i=1}^{D}Rp_i+RK\right)\right\}^{1/2}\right)\\ +\mathcal{O}_{\mathrm{p}}\left(\left\{\frac{\sum_{r=1}^{R_0}\|\mathrm{vec}(\mathbf{B}_{0r})\|_1}{s}\right\}\frac{n\sqrt{h_n}}{K^\tau}\right).\end{aligned} \tag{A.16}$$

It follows from (A.11) and (A.16) that

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

Further, by Assumption 1 and (A.28) of Lemma A.2, we have

$$\|\hat{m}(\mathbf{X})-m_0(\mathbf{X})\|^2 \le C_5 h_n\frac{1}{s^2}\|\hat{\mathbf{A}}^\flat-\mathbf{A}_0^\flat\|_{HS}^2 = C_5 h_n\frac{1}{s^2}\|\mathbf{a}^\sharp\|_2^2, \tag{A.17}$$

where C_5 is a constant, which will complete the proof of (14) by noting (A.5).

A.3 Proof of Theorem 2

Suppose $(\hat{\mathbf{A}}_{\mathrm{pen}}, \hat{\nu}_{\mathrm{pen}})$ is a solution to (10) and

$$\hat{\mathbf{A}}_{\mathrm{pen}} = \sum_{r=1}^{R}\hat{\boldsymbol{\beta}}_{r,1}\circ\hat{\boldsymbol{\beta}}_{r,2}\circ\cdots\circ\hat{\boldsymbol{\beta}}_{r,D}\circ\hat{\boldsymbol{\alpha}}_r.$$

Since the arguments used in the proof of Theorem 1 have probability inequality versions, we can show the consistency of the penalized estimator similarly. In the following, we let

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

where $\hat{\beta}_{r,d,\ell}$ is the ℓ-th element of $\hat{\boldsymbol{\beta}}_{r,d}$. By Theorem 3, there exists $\check{\nu}_{\mathrm{pen}} \in \mathbb{R}$ and

$$\check{\mathbf{A}}_{\mathrm{pen}} = \sum_{r=1}^{R}\hat{\boldsymbol{\beta}}_{r,1}\circ\hat{\boldsymbol{\beta}}_{r,2}\circ\cdots\circ\hat{\boldsymbol{\beta}}_{r,D}\circ\check{\boldsymbol{\alpha}}_r \in \mathbb{R}^{p_1\times\ldots\times p_D\times K},$$

such that

$$\check{\nu}_{\mathrm{pen}}+\frac{1}{s}\langle\check{\mathbf{A}}_{\mathrm{pen}},\Phi(\mathbf{X})\rangle = \hat{\nu}_{\mathrm{pen}}+\frac{1}{s}\langle\hat{\mathbf{A}}_{\mathrm{pen}},\tilde{\Phi}(\mathbf{X})\rangle, \tag{A.18}$$

where $\check{\boldsymbol{\alpha}}_r = (\check{\alpha}_{r,1},\ldots,\check{\alpha}_{r,K})^\intercal$ satisfying

$$\sum_{k=1}^{K}\check{\alpha}_{r,k}u_k = 0, \tag{A.19}$$

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

$$\mathbf{A}_0 = \sum_{r=1}^{R_0} \boldsymbol{\beta}_{0r,1} \circ \boldsymbol{\beta}_{0r,2} \circ \cdots \circ \boldsymbol{\beta}_{0r,D} \circ \boldsymbol{\alpha}_{0r}, \quad \boldsymbol{\alpha}_{0r} = (\alpha_{0r,1}, \ldots, \alpha_{0r,K})^\intercal, \quad \sum_{k=1}^{K} \alpha_{0r,k} u_k = 0.$$

The proof of Theorem 3 also shows that there exists $\tilde{\nu}_0 \in \mathbb{R}$ and

$$\tilde{\mathbf{A}}_0 = \sum_{r=1}^{R_0} \boldsymbol{\beta}_{0r,1} \circ \boldsymbol{\beta}_{0r,2} \circ \cdots \circ \boldsymbol{\beta}_{0r,D} \circ \tilde{\boldsymbol{\alpha}}_{0r} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times (K-1)}.$$

such that

$$\tilde{\nu}_0 + \frac{1}{s}\big\langle \tilde{\mathbf{A}}_0, \tilde{\Phi}(\mathbf{X}) \big\rangle = \nu_0 + \frac{1}{s}\big\langle \mathbf{A}_0, \Phi(\mathbf{X}) \big\rangle. \tag{A.20}$$

To achieve the norm 1 restrictions of (10), we normalize $\tilde{\boldsymbol{\alpha}}_{0r}$ in $\tilde{\mathbf{A}}_0$ by

$$\tilde{\mathbf{A}}_0 = \sum_{r=1}^{R_0} (\|\tilde{\boldsymbol{\alpha}}_{0r}\|_2 \cdot \boldsymbol{\beta}_{0r,1}) \circ \boldsymbol{\beta}_{0r,2} \circ \cdots \circ \boldsymbol{\beta}_{0r,D} \circ \frac{\tilde{\boldsymbol{\alpha}}_{0r}}{\|\tilde{\boldsymbol{\alpha}}_{0r}\|_2}.$$

Using the rescaling strategy (11) on $\{\|\tilde{\boldsymbol{\alpha}}_{0r}\|_2 \boldsymbol{\beta}_{0r,1}, \boldsymbol{\beta}_{0r,2}, \ldots, \boldsymbol{\beta}_{0r,D}\}$ for $r = 1, \ldots, R$, we get a solution $\{\rho_{0r,d}\}_{d=1}^{D}$, $r = 1, \ldots, R$. Denoting $\tilde{\boldsymbol{\beta}}_{0r,1} = \rho_{0r,1}\|\tilde{\boldsymbol{\alpha}}_{0r}\|_2 \boldsymbol{\beta}_{0r,1}$, $\tilde{\boldsymbol{\beta}}_{0r,d} = \rho_{0r,d}\boldsymbol{\beta}_{0r,d}$, $d = 2, \ldots, D$, we have

$$\tilde{\mathbf{A}}_0 = \sum_{r=1}^{R_0} \tilde{\boldsymbol{\beta}}_{0r,1} \circ \tilde{\boldsymbol{\beta}}_{0r,2} \circ \cdots \circ \tilde{\boldsymbol{\beta}}_{0r,D} \circ \frac{\tilde{\boldsymbol{\alpha}}_{0r}}{\|\tilde{\boldsymbol{\alpha}}_{0r}\|_2}.$$

Let

$$G_0 = \sum_{r=1}^{R_0} \sum_{d=1}^{D} \sum_{\ell=1}^{p_d} P_{\boldsymbol{\lambda}}(\tilde{\beta}_{0r,d,\ell}), \tag{A.21}$$

where $\tilde{\beta}_{0r,d,\ell}$ is the ℓ-th element of $\tilde{\boldsymbol{\beta}}_{0r,d}$, then it is the minimal value of penalty term on the corresponding scale class. Using (A.18)–(A.21), we can obtain

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

Let $\check{\mathbf{A}}_{\text{pen}}^{\flat} = \Omega(\check{\mathbf{A}}_{\text{pen}}, \check{\nu}_{\text{pen}})$ and $\mathbf{A}_0^{\flat} = \Omega(\mathbf{A}_0, \nu_0)$. Since $\hat{G} \geq 0$, we have

$$\sum_{i=1}^{n} \left(y_i - \frac{1}{s}\langle \check{\mathbf{A}}_{\text{pen}}^{\flat}, \Phi(\mathbf{X}_i) \rangle \right)^2 \leq \sum_{i=1}^{n} \left(y_i - \frac{1}{s}\langle \mathbf{A}_0^{\flat}, \Phi(\mathbf{X}_i) \rangle \right)^2 + G_0. \tag{A.22}$$

Similar to the proof of Theorem 1, we let $\mathbf{A}_{\text{pen}}^{\sharp} = \check{\mathbf{A}}_{\text{pen}}^{\flat} - \mathbf{A}_0^{\flat}$, $\mathbf{a}_{\text{pen}}^{\sharp} = \text{vec}(\mathbf{A}_{\text{pen}}^{\sharp})$, $\mathbf{a}_0^{\flat} = \text{vec}(\mathbf{A}_0^{\flat})$ and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\intercal \in \mathbb{R}^{n \times sK}$, where $\mathbf{z}_i = \text{vec}\{\Phi(\mathbf{X}_i)\}$, $i = 1, \ldots, n$. Using (A.22) and working out the squares, we obtain

$$\frac{1}{s^2}\|\mathbf{Z}\mathbf{a}_{\text{pen}}^{\sharp}\|_2^2 \leq 2\left\langle \frac{1}{s}\mathbf{Z}\mathbf{a}_{\text{pen}}^{\sharp}, \boldsymbol{\epsilon} \right\rangle + 2\left\langle \frac{1}{s}\mathbf{Z}\mathbf{a}_{\text{pen}}^{\sharp}, \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^{\flat} \right\rangle + G_0, \tag{A.23}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^\intercal$. By (A.19), Lemma A.1 and Lemma A.6, we have $\sum_{k=1}^{K} A^\sharp_{\boldsymbol{j},k} u_k = 0$ for $\boldsymbol{j} \in \mathcal{J}/\{(1, \cdots, 1)\}$. Since $\text{rank}(\mathbf{A}_0^\flat) \leq R_0 + 1$, $\text{rank}(\check{\mathbf{A}}^\flat_{\text{pen}}) \leq R + 1$, it is trivial to see $\text{rank}(\mathbf{A}^\sharp_{\text{pen}}) \leq R_0 + R + 2$. 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 $\|\mathbf{a}^\sharp_{\text{pen}}\|_2$ in (A.23).

Firstly, we will find the upper bound of $\langle \mathbf{Z}\mathbf{a}^\sharp_{\text{pen}}, \boldsymbol{\epsilon} \rangle$. For simplicity, we will use part of arguments in the proof of Theorem 1 with a slight abuse of notations. 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 $C > 0$. By (A.9)–(A.12) and Lemma A.5, we have the following upper bound

$$\langle \mathbf{Z}\mathbf{a}^\sharp_{\text{pen}}, \boldsymbol{\epsilon} \rangle \leq C\|\mathbf{a}^\sharp_{\text{pen}}\|_2 \left\{ nh_n \left(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK \right) \right\}^{1/2}, \tag{A.24}$$

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 $\langle \mathbf{Z}\mathbf{a}^\sharp_{\text{pen}}, \mathbf{y} - \boldsymbol{\epsilon} - \mathbf{Z}\mathbf{a}_0^\flat \rangle$. Using the Cauchy-Schwarz inequality, (A.11), (A.12), and (A.14), it shows that

$$\begin{aligned} \left\langle \frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp_{\text{pen}}, \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\rangle &\leq \left\| \mathbf{y} - \boldsymbol{\epsilon} - \frac{1}{s}\mathbf{Z}\mathbf{a}_0^\flat \right\|_2 \left\| \frac{1}{s}\mathbf{Z}\mathbf{a}^\sharp_{\text{pen}} \right\|_2 \\ &\leq \frac{C}{s}\|\mathbf{Z}\mathbf{a}^\sharp_{\text{pen}}\|_2 \left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{\sqrt{n}}{K^\tau} \\ &\leq \frac{C}{s}\|\mathbf{a}^\sharp_{\text{pen}}\|_2 \left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{n\sqrt{h_n}}{K^\tau}, \end{aligned} \tag{A.25}$$

with probability at least

$$1 - C_1 \exp\left\{ - C_2 \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right) \right\}.$$

Thirdly, applying (A.24) and (A.25) to (A.23), we get

$$\frac{1}{s^2}\|\mathbf{a}^\sharp_{\text{pen}}\|_2^2 \leq \frac{\delta_3}{s}\|\mathbf{a}^\sharp_{\text{pen}}\|_2 + \frac{1}{nh_n}G_0, \tag{A.26}$$

with probability at least

$$1 - C \exp\left\{ - C_5 \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right) \right\}, \tag{A.27}$$

where

$$\delta_3 = C\left\{ \frac{K\left(R^{D+1} + \sum_{i=1}^{D} Rp_i + RK\right)}{n} \right\}^{1/2} + C_6 \left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(\mathbf{B}_{0r})\|_1}{s} \right\} \frac{1}{K^{\tau - 1/2}}.$$

By solving the second order inequality (A.26), we obtain

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

under the same probability (A.27). To prove (15), we use the similar arguments of (A.17) to obtain,

$$\|\hat{m}_{\text{pen}}(\mathbf{X}) - m_0(\mathbf{X})\|^2 \leq \frac{C\{\delta_3^2 + (4KG_0)/n\}}{K},$$

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.4 Technical results

Lemma A.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 $\mathcal{J}$ *is defined in* (A.1).

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 A.2. *Suppose* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$ *and* $\mathbf{U} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ *is a random tensor with its entry* $U_{\boldsymbol{j}} \overset{i.i.d.}{\sim} U(0,1)$*, for* $\boldsymbol{j} \in \mathcal{J}$*, where* $\mathcal{J}$ *is defined in* (A.1)*. Recall that* $(\Phi(\mathbf{X}))_{\boldsymbol{j},k} = b_k(X_{\boldsymbol{j}})$*, where* $\{b_k(x)\}_{k=1}^K$ *be a B-spline basis,* $x \in [0,1]$*. Under Assumptions 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.

$$C_1 C_\zeta h_n \|\mathbf{A}\|_{HS}^2 \le \mathbb{E}\{\langle \mathbf{A}, \Phi(\mathbf{X})\rangle^2\} \le C_2 h_n \|\mathbf{A}\|_{HS}^2, \tag{A.28}$$

and

ii.

$$\|\langle \mathbf{A}, \Phi(\mathbf{X})\rangle\|_{\psi_2}^2 \le C_3 \tilde{h}_n \|\mathbf{A}\|_{HS}^2, \tag{A.29}$$

where C_1, C_2, C_3, C_ζ *are positive constants,* C_ζ *depends on the order of B-spline* ζ*, and* $\tilde{h}_n$ *is defined in* (A.4)*.*

Proof. We will prove the population bound (A.28) 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 $1 \le q \le +\infty$,

$$C_\zeta \|\mathbf{A}_{\boldsymbol{j}}\|_q \le h_n^{-1/q} \left\| \sum_{k=1}^K A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}}) \right\|_q \le C \|\mathbf{A}_{\boldsymbol{j}}\|_q, \tag{A.30}$$

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.30) yields

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

then

$$C_\zeta h_n \|\mathbf{A}\|_{HS}^2 \le \mathbb{E}\{\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2\} \le C h_n \|\mathbf{A}\|_{HS}^2. \tag{A.31}$$

By Assumption 1, we have

$$C_1 \mathbb{E}\{\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2\} \le \mathbb{E}\{\langle \mathbf{A}, \Phi(\mathbf{X})\rangle^2\} \le C_4 \mathbb{E}\{\langle \mathbf{A}, \Phi(\mathbf{U})\rangle^2\}. \tag{A.32}$$

It follows from (A.31) and (A.32) that

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

which completes the proof of (A.28).

Now, we will prove the sub-Gaussian norm bound (A.29). Note that

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

then

$$\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \le 2\Big\| \sum_{\boldsymbol{j}\in\mathcal{J}_1}\sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2 + 2\Big\| \sum_{k=1}^{K} A_{1,\cdots,1,k} b_k(U_{1,\cdots,1})\Big\|_{\psi_2}^2. \tag{A.33}$$

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

$$\Big\| \sum_{\boldsymbol{j}\in\mathcal{J}_1}\sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2 \le C_5 \sum_{\boldsymbol{j}\in\mathcal{J}_1}\Big\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2. \tag{A.34}$$

It follows from (A.33) and (A.34) that

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

Therefore,

$$\begin{aligned}
&\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \\
&\le (2C_5+2)\sum_{\boldsymbol{j}\in\mathcal{J}}\Big\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2 + (2C_5+2)\Big\| \sum_{k=1}^{K} A_{1,\cdots,1,k} b_k(U_{1,\cdots,1})\Big\|_{\psi_2}^2 \\
&= (2C_5+2)\sum_{\boldsymbol{j}\in\mathcal{J}}\Big\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_{\psi_2}^2.
\end{aligned} \tag{A.35}$$

We then consider the sub-Gaussian norm of $\sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})$. When $q = 1$, by (A.30), we have

$$\Big\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\Big\|_1 \le 2\frac{\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\|_2}{\sqrt{2}} \le C\sqrt{h_n}\|\mathbf{A}_{\boldsymbol{j}}\|_2. \tag{A.36}$$

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

$$\frac{\| \sum_{k=1}^{K} A_{\boldsymbol{j},k} b_k(U_{\boldsymbol{j}})\|_q}{\sqrt{q}} \le C\frac{h_n^{1/q}}{\sqrt{q}}\|\mathbf{A}_{\boldsymbol{j}}\|_q \le C\frac{h_n^{1/q}}{\sqrt{q}}\|\mathbf{A}_{\boldsymbol{j}}\|_2. \tag{A.37}$$

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

$$\frac{h_n^{1/q}}{\sqrt{q}}\|\mathbf{A}_{\boldsymbol{j}}\|_2 \le \frac{h_n^{1/(-2\log h_n)}}{(-2\log h_n)^{1/2}}\|\mathbf{A}_{\boldsymbol{j}}\|_2. \tag{A.38}$$

Due to the definition of $\tilde{h}_n$ in (A.4) and using (A.35)–(A.38), we get

$$\|\langle \mathbf{A}, \Phi(\mathbf{U})\rangle\|_{\psi_2}^2 \le (2C_5+2)\tilde{h}_n C^2 \|\mathbf{A}\|_{HS}^2. \tag{A.39}$$

Note that for $q \ge 1$,

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

therefore,

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

which completes the proof of (A.29). □

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

$$\mathcal{P} = \left\{ \frac{\mathrm{vec}(\mathbf{A})}{\|\mathbf{A}\|_{HS}} : \sum_{k=1}^{K} A_{\boldsymbol{j},k} u_k = 0, \text{ for } \boldsymbol{j} \in \mathcal{J}/\{(1,\ldots,1)\}, \ \mathrm{rank}(\mathbf{A}) \le R \right\}, \tag{A.40}$$

where $u_k = \int_0^1 b_k(x)\mathrm{d}x$ *and* $\mathcal{J}$ *is defined in* (A.1). *The Gaussian width satisfying*

$$w(\mathcal{P}) \le C\left(R^{D+1} + R\sum_{d=1}^{D} p_d + RK \right)^{1/2}. \tag{A.41}$$

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

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

where $C_1 = 3D + 4$ is a constant. Suppose $\mathbf{a} \in \mathcal{P}$ and $\mathbf{x} \in \mathcal{N}(\mathbf{0}, \mathbf{I}_{s\times s})$, then by the Dudley's integral entropy bound (see, e.g., Theorem 3.1 of Koltchinskii, 2011), we obtain

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

Thus we complete the proof. □

Lemma A.4. *Let* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$ *and suppose* $\mathcal{P}$ *is defined in* (A.40). *Under Assumptions 1 and 4 , we have*

i.

$$\sup_{\mathrm{vec}(\mathbf{A})\in\mathcal{P}} \left| \frac{1}{n} \frac{1}{\mathbb{E}\{|\langle \mathbf{A}, \Phi(\mathbf{X})\rangle|^2\}} \sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i)\rangle^2 - 1 \right| \le C_1 \tilde{h}_n h_n^{-1} \frac{w(\mathcal{P})}{\sqrt{n}} \tag{A.42}$$

with probability at least $1 - \exp\{-C_2 w^2(\mathcal{P})\}$, *where* $w(\mathcal{P})$ *is the Gaussian width,* $(\Phi(\mathbf{X}))_{\boldsymbol{j},k} = b_k(X_{\boldsymbol{j}})$ *for* $\boldsymbol{j} \in \mathcal{J}$, $k = 1, \ldots, K$, $\mathcal{J}$ *is defined in* (A.1) *and* $\tilde{h}_n$ *is defined in* (A.4). *Furthermore, suppose* $n > C\tilde{h}_n^2 h_n^{-2} w^2(\mathcal{P})$ *for some* $C > 0$, *then with the same probability, we have*

ii.

$$C_3 h_n \le \inf_{\mathrm{vec}(\mathbf{A})\in\mathcal{P}} \frac{1}{n} \left| \sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i)\rangle \right|^2 \le \sup_{\mathrm{vec}(\mathbf{A})\in\mathcal{P}} \frac{1}{n} \left| \sum_{i=1}^{n} \langle \mathbf{A}, \Phi(\mathbf{X}_i)\rangle \right|^2 \le C_4 h_n. \tag{A.43}$$

Note that the above $w(\mathcal{P})$ *can be replaced by a constant* t, *provided* $t \ge w(\mathcal{P})$.

Proof. Based on Lemma A.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,

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

and by Lemma A.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 $1 - \exp(-C_8 \theta^2 n / \kappa_n^4)$, we have (A.42) holds with $C_1 = C_7 C_6 C_5^2$ and $C_2 = C_8 C_7^2 C_6^2$. Suppose $\sqrt{n} > C \tilde{h}_n h_n^{-1} w(\mathcal{P})$ for some $C > 0$, then by Lemma A.2, with probability at least $1 - \exp\{-C_2 w^2(\mathcal{P})\}$, we have

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

which completes the proof of (A.43). □

Lemma A.5. *Suppose* $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$, $\mathrm{rank}(\mathbf{A}) \leq R$ *and* $\sum_{k=1}^K A_{\boldsymbol{j},k} u_k = 0$ *for* $\boldsymbol{j} \in \mathcal{J}/\{(1, \ldots, 1)\}$, *where* $u_k = \int_0^1 b_k(x) \mathrm{d}x$ *and* $\mathcal{J}$ *is defined in* (A.1). *Under Assumptions 1, 2, and 4, 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* $C > 0$, *where* $\tilde{h}_n$ *is defined in* (A.4), *we then have*

$$\sum_{i=1}^n \langle \mathbf{A}, \Phi(\mathbf{X}_i)\rangle \epsilon_i \leq C_1 \|\mathbf{A}\|_{HS} \left\{ n h_n \left(R^{D+1} + \sum_{i=1}^D R p_i + RK \right) \right\}^{1/2}, \tag{A.44}$$

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 (A.7), then the left hand side of (A.44) 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

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

where $\lambda_{\mathrm{Rmax}}(\mathbf{Z}^\intercal\mathbf{Z}) = \sup_{\mathbf{a}\in\mathcal{P}} \|\mathbf{Z}\mathbf{a}\|_2$ and $\mathcal{P}$ is defined as (A.40). By the covering number argument in Lemma A.7,

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

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

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

By Assumption 2, for $\boldsymbol{\eta} \in \Gamma_1$, $\mathbb{E}\{\exp(t\boldsymbol{\eta}^\intercal\boldsymbol{\epsilon})\} \leq \exp(Ct^2\|\boldsymbol{\eta}\|^2) \leq \exp(Ct^2)$. 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

$$(\mathbf{Z}\mathbf{a})^\intercal\boldsymbol{\epsilon} \leq C_7 \sqrt{\lambda_{\mathrm{Rmax}}(\mathbf{Z}^\intercal\mathbf{Z})} \left(R^{D+1} + R\sum_{i=1}^{D} p_i + RK \right)^{1/2}, \tag{A.45}$$

with probability as least

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

Plugging (A.41) and (A.43) into (A.45), we will complete the proof of (A.44). □

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

$$N(\epsilon, \Gamma_2, l_2) \leq \left(\frac{3D+4}{\epsilon} \right)^{R^{D+1} + R\sum_{d=1}^{D+1} p_d}. \tag{A.46}$$

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

$$\mathbf{A} = \mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D \times_{D+1} \mathbf{B}_{D+1}, \tag{A.47}$$

where $\mathbf{I} \in \mathbb{R}^{R \times R \ldots \times R}$ is a diagonal tensor of which all the diagonal entries are 1, $\mathbf{B}_d \in \mathbb{R}^{p_d \times R}$, and $\times_d$ denotes the d-mode (matrix) product of a tensor with a matrix (Kolda and Bader, 2009). Let $r_d = \mathrm{rank}(\mathbf{B}_d)$. Through the QR decomposition, we get $\mathbf{B}_d = \mathbf{Q}_d \mathbf{R}_d$, where $\mathbf{Q}_d^\intercal \mathbf{Q}_d = \mathbf{I}_{r_d}$ with $\mathbf{I}_{r_d} \in \mathbb{R}^{r_d \times r_d}$ the identity matrix. Applying the argument to (A.47), 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

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

Therefore (A.46) is shown by noting that $r_d \leq R$ for $d = 1, \ldots, D+1$. □

B Algorithmic analysis

B.1 Equivalent basis

To begin with, we define or recall some notations which will be used later. Recall that $\{\tilde{b}_k(x)\}_{k=1}^K$ 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 $\tilde{u}_k = \int_0^1 \tilde{b}_k(x)\mathrm{d}x$. Denote $\Phi(\mathbf{X}), \check{\Phi}(\mathbf{X}) \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D \times K}$ be the tensor formed from the bases, which means $(\Phi(\mathbf{X}))_{\boldsymbol{j},k} = b_k(X_{\boldsymbol{j}})$ and $(\check{\Phi}(\mathbf{X}))_{\boldsymbol{j},k} = \tilde{b}_k(X_{\boldsymbol{j}})$, $\boldsymbol{j} \in \mathcal{J}$, $k = 1, \ldots, K$. We define two function classes,

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

and

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

where $\nu_l \in \mathbb{R}$, $\boldsymbol{\beta}_{lr,d} \in \mathbb{R}^{p_d}$ and $\boldsymbol{\alpha}_{lr} = (\alpha_{lr,1}, \cdots, \alpha_{lr,K})^\intercal \in \mathbb{R}^K$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, $l = 1, 2$. Recall that $\tilde{\Phi}(\mathbf{X}) \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K-1}$ is defined by $(\tilde{\Phi}(\mathbf{X}))_{\boldsymbol{j},k} = \tilde{b}_{k+1}(X_{\boldsymbol{j}})$, $\boldsymbol{j} \in \mathcal{J}$, $k = 1, \ldots, K-1$. We define the following function class that the linear constraints are removed, 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 3, we can remove the linear constraints in (7) and use any equivalent spline basis to develop our theory.

Theorem 3. $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}_3$.

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 $\mathbf{b}(x) = \mathbf{Q}\tilde{\mathbf{b}}(x)$, where $\mathbf{b}(x) = (b_1(x), \ldots, b_K(x))^\intercal$ and $\tilde{\mathbf{b}}(x) = (\tilde{b}_1(x), \ldots, \tilde{b}_K(x))^\intercal$. 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

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

and $\mathbf{J} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ as the tensor of which all the entries are 1. 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 $\sum_{k=1}^{K-1} \alpha_{3r,k}\tilde{u}_{k+1} = C_r$, it is trivial to see $\tilde{u}_1 \neq 0$. We can choose $\alpha_{2r,1} = -C_r/\tilde{u}_1$, $\alpha_{2r,k+1} = \alpha_{3r,k}$ for $k = 1, \ldots, K-1$ so that $\boldsymbol{\alpha}_{2r}$ satisfies the constraint in $\mathcal{M}_2$. Taking $\nu_2 = \nu_3 + \sum_{r=1}^{R} \langle \mathbf{B}_{3r}, C_r/\tilde{u}_1 \mathbf{J} \rangle$, $\mathbf{B}_{2r} = \mathbf{B}_{3r}$, 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 $\mathcal{M}_3 = \mathcal{M}_2$.

□

B.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 (11) becomes

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

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

B.3 Proof of Proposition 1

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 $\prod_d \rho_{r,d} = 1$, $\rho_{r,d} > 0$ 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 (B.1) 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.

B.4 Proof of Proposition 2

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.

C Identifiability

It is noted that our theory does not require the identifiability for each component in (3). For completeness, we discuss the following identifiable problems. To begin with, we state the uniqueness of the representation (3), which means that (3) 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.,

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

   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 (3), 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 (3).

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 (3). 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 (3) 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 $\mathrm{Span}\{\tilde{f}_r\}_{r=1}^R \subsetneq \mathrm{Span}\{f_r\}_{r=1}^R$. We say the representation is identifiable if the components are unique up to permutation and scaling. 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 (3). We then list some sufficient conditions to achieve the identifiability, based on the fundamental idea of the identifiability for CP decomposition. Denote

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

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 (3) 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 (3) 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 (3) 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 (3) is unique up to permutation and scaling.

For simplicity, we present the general condition in the following theorem. In the proof of Theorem 4, we in fact prove all the aforementioned sufficient conditions.

Theorem 4 (Identifiability)**.** *If*

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

then the representation (3) *is unique up to permutation and scaling.*

Proof. Suppose there is another representation of (3), i.e,

$$\begin{aligned} m(\mathbf{X}) &= \nu + \frac{1}{s}\sum_{r=1}^{R} \langle \boldsymbol{\beta}_{r,1} \circ \boldsymbol{\beta}_{r,2} \circ \cdots \circ \boldsymbol{\beta}_{r,D}, F_r(\mathbf{X}) \rangle \\ &= \bar{\nu} + \frac{1}{s}\sum_{r=1}^{R} \left\langle \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D}, \bar{F}_r(\mathbf{X}) \right\rangle, \end{aligned} \tag{B.3}$$

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 (B.3). In the remaining sum of inner products, we consider the following arguments. Suppose the minimal bases of the vector space

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

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

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

and

$$\bar{\mathbf{A}}^f = \frac{1}{s}\sum_{r=1}^{R} \bar{\boldsymbol{\beta}}_{r,1} \circ \bar{\boldsymbol{\beta}}_{r,2} \circ \cdots \circ \bar{\boldsymbol{\beta}}_{r,D} \circ \bar{\boldsymbol{\eta}}_r, \tag{B.5}$$

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 (B.3) equals, i.e.,

$$\langle \mathbf{A}^f, \Psi(\mathbf{X}) \rangle = \langle \bar{\mathbf{A}}^f, \bar{\Psi}(\mathbf{X}) \rangle. \tag{B.6}$$

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

$$\text{Span}\{\psi_{k^\star}(x)\}_{k^\star=1}^{K^\star} = \text{Span}\{\bar{\psi}_{\bar{k}^\star}(x)\}_{\bar{k}^\star=1}^{\bar{K}^\star}. \tag{B.7}$$

Based on (B.7), 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 (B.6) as

$$\langle \mathbf{A}^f, \Psi(\mathbf{X}) \rangle = \langle \bar{\mathbf{A}}^f, \Psi(\mathbf{X}) \rangle. \tag{B.8}$$

Secondly, we will show $\mathbf{A}^f = \bar{\mathbf{A}}^f$ in (B.8). In the end, we will take the advantages of identifiable theory about CP decomposition and complete the proof.

To show (B.7), 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 A.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 $\text{Span}\{\tilde{f}_r\}_{r=1}^R \subsetneq \text{Span}\{f_r\}_{r=1}^R$, such the representation (3) 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 (B.7) holds and $\bar{K}^\star = K^\star$.

To show $\mathbf{A}^f = \bar{\mathbf{A}}^f$ in (B.8), we let $\mathbf{A}^{f,\star} = \mathbf{A}^f - \bar{\mathbf{A}}^f$. It 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 denote the corresponding value

$$C_{-\boldsymbol{j}_0} = \sum_{\boldsymbol{j}\neq\boldsymbol{j}_0} \sum_{k^\star=1}^{K^\star} A^{f,\star}_{\boldsymbol{j},k^\star} f_{k^\star}(X_{\boldsymbol{j}}).$$

It then shows that

$$\sum_{k^\star=1}^{K^\star} A^{f,\star}_{\boldsymbol{j}_0,k^\star} \psi_{k^\star}(X_{\boldsymbol{j}_0}) + C_{-\boldsymbol{j}_0} = 0, \tag{B.9}$$

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 A.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 (B.9) 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, (B.4) 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 (3) 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 (B.5) and the decomposition (B.4) are the same up to permutation and scaling. Therefore the representation (3) is unique up to permutation and scaling. Now, to make the representation (3) 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 write $\mathbf{B}_{D+1} := \boldsymbol{\eta} = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_R)$ 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 (B.4) 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 (B.4) 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 (B.4) 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 (3) 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 (3) 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 (3) 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 (3) is unique up to permutation and scaling.

Since $\{f_r\}_{r=1}^R$ are allowed to be the same in the model, we present 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). □

D Starting points

Motivated by the initial point strategy used in the MATLAB toolbox TensorReg, we propose a sequential down-sizing strategy. For the penalized tensor linear regression, TensorReg applies the unpenalized tensor linear regression on a down-sized sample first. The down-sized sample depends on a shrinkage parameter $\vartheta = n/(CR\sum_{d=1}^D p_d)$, where C is a constant supplied by users. If $\vartheta \leq 1$, the down-sized sample is just the original sample $(\mathbf{X}_i, y_i)$; if not, $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is down-sized to a smaller tensor of size $\tilde{p}_1 \times \ldots \times \tilde{p}_D$, where $\tilde{p}_d = \lfloor p_d/\vartheta \rfloor$. Secondly, transform the solution of coefficient tensor of the previous unpenalized method back to the original size, and then run the penalized algorithm.

Our sequential down-sizing procedure refers to considering a sequential down-size $\tilde{p}_1^{(1)} \times \cdots \times \tilde{p}_D^{(1)} \times K$, $\tilde{p}_1^{(2)} \times \cdots \times \tilde{p}_D^{(2)} \times K$, $\cdots$, $\tilde{p}_1^{(\eta)} \times \cdots \times \tilde{p}_D^{(\eta)} \times K$ of the samples in the initial stage. Here we can choose $\{p_d^{(t)}\}_{t=1}^{\eta} \subset (C, p_d) \cap \mathbb{N}$ as an arithmetic sequence, where C is a constant defined by users. Firstly, we use random initial points for the unpenalized tensor regression on the down-size sample with size of $\tilde{p}_1^{(1)} \times \cdots \times \tilde{p}_D^{(1)} \times K$, and then use the results of the unpenalized tensor regression under the size of $\tilde{p}_1^{(i)} \times \cdots \times \tilde{p}_D^{(i)} \times K$ as the initial points (after up-size) for that of $\tilde{p}_1^{(i+1)} \times \ldots \times \tilde{p}_D^{(i+1)} \times K$, where $1 \leq i < \eta$. Finally, use the results (after

up-size) under the size of $\tilde{p}_1^{(\eta)} \times \cdots \times \tilde{p}_D^{(\eta)} \times K$ as the final initial points for the penalized method.

In the grids search procedure, when R and λ_2 are fixed, we apply the sequential down-sizing initial strategy to the smallest λ_1 in the grids, and then use the results as the initial points of the second smallest λ_1 in grids. Next, we use the new results as the initial points of the third smallest λ_1 and repeat this procedure for all values of λ_1.

E Grids of tuning parameters

In synthetic experiments, we considered the following grids: $R \in \{1, 2, 3, 4, 5\}$, $\lambda_1 \in \{10^{-2}, 5 \times 10^{-1}, 10^{-1}, \ldots, 10^2, 5 \times 10^2, 10^3\}$ and $\lambda_2 \in \{0, 0.5, 1\}$, for BroadcasTR, TLR, and TLR-rescaled. For ENetR, we used the same grids of λ_1 and λ_2. For BroadcasTR, similar to Huang et al. (2010), we used the cubic spline and fixed the number of basis $K = 7$. The knots were chosen as the equally spaced quantiles.

In the real data analysis, we considered the following grids for the rank $R \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, and penalized parameters $\lambda_1 \in \{10^{-2}, 2.5 \times 10^{-2}, 5 \times 10^{-2}, 7.5 \times 10^{-2}, 10^{-1}, \ldots, 10^2, 2.5 \times 10^2, 5 \times 10^2, 7.5 \times 10^2, 10^3\}$, $\lambda_2 \in \{0, 0.5, 1\}$ for BroadcasTR, TLR, TLR-rescaled, and the same grids of λ_1 and λ_2 for ENetR. We used the same spline basis setting as in the previous synthetic experiments for BroadcasTR.

F Data preprocessing of monkey's electrocorticography data

Our data preprocessing procedure is similar to Chao et al. (2010) and Shimoda et al. (2012). Firstly, the original signals were band-pass filtered from 0.3 to 499Hz and re-referenced using a common average reference montage (McFarland et al., 1997). Secondly, we use Morlet wavelet transformation to get the time-frequency representation at time t, where there are ten different center frequencies (20Hz, 30Hz, ..., 110 Hz) and ten time lags $(t - 900\,\text{ms}, t - 800\,\text{ms}, \ldots, t - 100\,\text{ms}, t)$. Finally, after a standardization step (z-score) at each frequency over the 10 time lags for each electrode, we get our input tensor of size $64 \times 10 \times 10$, such that the values of each entry lie in $\mathcal{I} = [-2.75, 2.85]$.

References

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*.

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.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1.

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

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

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.

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.

McFarland, D. J., McCane, L. M., David, S. V. and Wolpaw, J. R. (1997) Spatial filter selection for eeg-based communication. *Electroencephalography and Clinical Neurophysiology*, **103**, 386–394.

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 predictors. *Biometrics*, **66**, 61–69.

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

Wood, S. N. (2017) *Generalized additive models: an introduction with R*. Boca Raton: Chapman and Hall/CRC.

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

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.