More Efficient Estimation of Multivariate Additive Models Based on Tensor Decomposition and Penalization

Liu, Xu; Lian, Heng; Huang, Jian

Published in:
Journal of Machine Learning Research

Published: 01/01/2024

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Publication details:
Liu, X., Lian, H., & Huang, J. (2024). More Efficient Estimation of Multivariate Additive Models Based on Tensor Decomposition and Penalization. *Journal of Machine Learning Research*, *25*, Article 161.

Journal of Machine Learning Research 25 (2024) 1-27 Submitted 5/22; Revised 4/23; Published 5/24

More Efficient Estimation of Multivariate Additive Models Based on Tensor Decomposition and Penalization

Xu Liu LIU.XU@SUFE.EDU.CN
School of Statistics and Management
Shanghai University of Finance and Economics, Shanghai, China
and
Yunnan Key Laboratory of Statistical Modeling and Data Analysis
Yunnan University, Kunming, China

Heng Lian HENGLIAN@CITYU.EDU.HK
City University of Hong Kong Shenzhen Research Institute
Shenzhen, China
and
Department of Mathematics
City University of Hong Kong, Kowloon Tong, Hong Kong, China

Jian Huang J.HUANG@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University, Hong Kong, China

Editor: Animashree Anandkumar

Abstract

We consider parsimonious modeling of high-dimensional multivariate additive models using regression splines, with or without sparsity assumptions. The approach is based on treating the coefficients in the spline expansions as a third-order tensor. Note the data does not have tensor predictors or tensor responses, which distinguishes our study from the existing ones. A Tucker decomposition is used to reduce the number of parameters in the tensor. We also combined the Tucker decomposition with penalization to enable variable selection. The proposed method can avoid the statistical inefficiency caused by estimating a large number of nonparametric functions. We provide sufficient conditions under which the proposed tensor-based estimators achieve the optimal rate of convergence for the nonparametric regression components. We conduct simulation studies to demonstrate the effectiveness of the proposed novel approach in fitting high-dimensional multivariate additive models and illustrate its application on a breast cancer copy number variation and gene expression data set.

Keywords: High dimensionality; Sparse models; Splines; Tensor estimation; Tucker decomposition.

1 Introduction

Linear regression is one of the oldest and the most popular statistical tools used for relating predictors to a continuous response. It imposes the strict assumption that the effect between any predictor and the response is linear, which may not be satisfied in some applications

(Liu et al., 2011). Semiparametric models such as those with additive structures also have a long history in statistics and have been actively investigated (Stone, 1985, 1986). The additive form in the conditional mean circumvents the curse of dimensionality problem in nonparametric regression. They are more parsimonious than fully nonparametric models that are difficult to fit when the number of predictors p is larger than three (Stone, 1980), and more flexible than linear models by not constraining the relationships to be linear.

In the multivariate additive regression problem, we assume the generating model

$$y_{il} = \mu_l + \sum_{j=1}^{p} f_{jl}(x_{ij}) + \epsilon_{il}, i = 1, \ldots, n, l = 1, \ldots, q, \tag{1}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})^{\mathrm{T}}$ are the q responses, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are the p predictors, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq})^{\mathrm{T}}$ are the mean zero errors for the q responses which can be correlated. The functions f_{jl} are used to model the relationship between the j-th predictor x_{ij} and the l-th response y_{il}. The goal is to estimate these unknown functions based on the sample. If we impose the constraint that f_{jl} is a linear function (with unknown linear coefficient), this reduces to the standard linear regression model. On the other hand, the fully nonparametric model is given by

$$y_{il} = \mu_l + \sum_{j=1}^{p} f_l(x_{i1,\ldots,x_{ip}}) + \epsilon_{il}, i = 1, \ldots, n, l = 1, \ldots, q,$$

which involves the estimation of q p-dimensional functions $f_l, l = 1, \ldots, q$. It is well-known that the sample size required to estimate a p-dimensional function for p large is excessive (Stone, 1980). Thus in practice nonparametric regression is rarely used for $p > 3$. More concretely, for univariate response ($q = 1$), if we use the spline estimator to estimate the functions and the number of basis functions for each dimension is K, for estimating a p-dimensional function, the number of parameters is K^p, while the number of parameters for additive models is Kp (He and Shi, 1996). Thus, for nonparametric regression, the number of parameters increases exponentially fast with dimension, and in this sense we say the semiparametric additive model is more parsimonious.

In this work, we allow both q and p to be diverging with the sample size. The additive model has a well-known identifiability issue in the sense that adding a scalar c to f_{jl} and subtracting the same c from $f_{j'l}$ for some $j' \neq j$ leads to the same regression function. In other words, without additional assumptions, f_{jl} can only be estimated up to an arbitrary constant. Thus we take up the standard identifiability assumption $\int f_{jl}(x)dx = 0$ which is commonly used (Liu et al., 2011; Fan et al., 2014). This constraint can be enforced easily for the spline estimation approach as we detail in Section 2.1.

Even in semiparametric models which already partially addressed the problem of curse of dimensionality to some extent, we may still have a large number of univariate functions to estimate, raising efficiency and stability concerns of the statistical estimation. One can take advantage of the widely adopted concept of sparsity to alleviate the problem, which works under the assumption that some covariates do not have effects on the responses. Thus it is desired to find and remove those covariates either for efficiency reasons or to make the model more easily interpretable. Such zero components in additive models can be found by

optimizing a penalized functional as in Xue (2009); Ravikumar et al. (2008); Meier et al. (2009); Huang et al. (2010); Wang et al. (2011), even in the ultra-high dimensional setting.

The assumption of sparsity has been shown to be valuable in the literature. A different scenario for additive models is that although all component functions are nonzero, the functions are similar or related in some sense so that one can borrow information among the functions when simultaneously estimating them. Motivated by this, we herein propose a different route to achieving further parsimony in multivariate additive models, which can also be combined with the working assumption of sparsity to take advantage of both worlds.

For fitting additive models, popular approaches include kernel methods and splines (Linton and Nielsen, 1995; Liu et al., 2011). However, kernel methods are hard to be adapted to the high-dimensional setting due to its high computational cost, as discussed in Liu et al. (2011). In the high-dimensional setting, all existing studies with implementation based on the series estimation method are based spline methods (Meier et al., 2009; Huang et al., 2010; Fan et al., 2014). Thus we focus on B-spline approach in this paper. After approximating all component functions using splines, we will see the spline coefficients naturally organize themselves into the form of a third-order tensor. Then the way to parsimonious modelling is to assume that the tensor has a low rank. Conceptually, low-rankness of tensors can be regarded as a higher-order extension of low-rankness of matrices, and the latter has become common in the statistics community for decades. In particular, reduced rank regression (Izenman, 1975; Geweke, 1996; Anderson, 1999; Bunea et al., 2011; Chen et al., 2013) is built upon the assumption that the coefficient matrix has a low rank in multivariate linear regression. When the covariates are matrices, which appear naturally in applications such as medical imaging, Negahban and Wainwright (2011); Zhou and Li (2014) used low-rankness assumption in such matrix regression problems. As a direct extension, when the covariates or responses or both are higher-order tensors, Zhou et al. (2013); Raskutti et al. (2017); Sun and Li (2017); Sun et al. (2017) used either sparsity or low-rankness or both assumptions for tensors. All these models concerned are parametric models.

Tensor estimation has received increasing attention recently in the statistical community, for example in Zhou et al. (2013); Sun et al. (2017); Miranda et al. (2015); Raskutti et al. (2017) to name a few among possibly many others. In all these works, parametric models are considered and tensor structure appears naturally due to that either the responses or the covariates are tensors. The current study is fundamentally different from those in that the model we study is semiparametric. The covariates and responses are random vectors as in all traditional statistical problems instead of being tensors. The tensor structure arises only after using the series estimation approach for fitting the semiparametric model. Thus our study reveals a new class of statistical problems in which tensor estimation can play a role. We also provide an implementation of the proposed method through a publicly available R package `tensorMAM` (`https://github.com/xliusufe/tensorMam`). Our main contributions are summarized as follows.

- First, methodologically, we proposed a new estimation approach for semiparametric additive models based on tensor decomposition, to reduce the number of parameters to be estimated. Although both tensor regression and additive models are widely studied already, this combination is novel in that we have identified a class of classical models that appears to have nothing to do with tensor regression and use tensor techniques to improve the estimation performance.

- Second, we establish the statistical theory for the proposed estimator, which clearly shows the role of the reduced number of parameters resulting in an improved convergence rate. When further combined with penalization, we establish the oracle property of the sparse estimator. In particular, to show the improved convergence rate, we need to use empirical processes techniques and Dudley's integral bound which were not necessary in previous works on high-dimensional additive models based on splines.

- Finally, we numerically compare the proposed estimator with existing methods for fitting high-dimensional additive models, including the traditional method without using low-rank factorization, and some more recent methods based on matrix decomposition (see also Section 2.3). We show that our method has clear advantages in some settings.

To be self-complete, in the last part of the introduction, we introduce some notation and operations involving tensors (Kolda and Bader, 2009). Suppose $\mathbf{T} = \{t_{i_1 \ldots i_N}\} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is a tensor of order N. The only norm of a tensor used in this paper is the squared root of the sum of squares of all its elements, i.e., $\|\mathbf{T}\| = \sqrt{\sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} |t_{i_1 \cdots i_N}|^2}$, which is called the Frobenius norm. A fiber of $\mathbf{T}$ is defined by fixing all indices but one, assumed to be oriented as column vectors. Slices are two-dimensional sections of a tensor, obtained by fixing all but two indices. Matricization is the process of reordering the elements of a tensor into a matrix. In particular, the mode-n matricization of $\mathbf{T}$, denoted by $\mathbf{T}_{(n)}$, arranges the mode-n fibers to be the columns of the resulting $I_n \times (I_1 \cdots I_{n-1} I_{n+1} \cdots I_N)$ matrix. More specifically, the tensor element $(i_1, i_2, \ldots, i_N)$ is mapped to matrix element (i_n, j), where $j = 1 + \sum_{k=1, k \neq n}^{N} (i_k - 1) J_k$ with $J_k = \sum_{m=1, m \neq n}^{N} I_m$. Tensors can be multiplied together, and in this paper we only use the mode-n product of a tensor with a matrix. For a matrix $\mathbf{U} = \{u_{ji}\}$ of size $J_n \times I_n$, the mode-n product of $\mathbf{T}$ and $\mathbf{U}$, denoted by $\mathbf{T} \times_n \mathbf{U}$ is a tensor of size $I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N$, with elements $(\mathbf{T} \times_n \mathbf{U})_{i_1 \cdots i_{n-1} j i_n \cdots i_N} = \sum_{i_n} t_{i_1 \cdots i_N} u_{j i_n}$. For later use, we mention the property that for n-th order tensors $\mathbf{S}$ and $\mathbf{T}$,

$$\mathbf{T} = \mathbf{S} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \cdots \times_N \mathbf{A}_n \Leftrightarrow \mathbf{T}_{(n)} = \mathbf{A}_n \mathbf{S}_{(n)} (\mathbf{A}_N \otimes \cdots \otimes \mathbf{A}_{n+1} \otimes \mathbf{A}_{n-1} \otimes \cdots \otimes \mathbf{A}_1)^{\mathrm{T}}, \quad (2)$$

where $\otimes$ denotes the Kronecker product of matrices.

The rest of the article is organized as follows. In Section 2 we present our methodology based on B-splines expansion and Tucker decomposition for tensors. We then discuss in detail the relationship between our approach and some other existing ones. We also establish the convergence rate of the estimator. In Section 3, variable selection based on sparsity assumption is further incorporated so that we can deal with even larger dimensions. Section 4 discusses the computational algorithm based on group coordinate descent. Section 5 reports our simulation studies and an empirical application. We conclude with some discussions in Section 6. The proofs are given in the Appendix. The Appendix also contains a table of notations for ease of reference.

2 Tensor estimation of multivariate additive models

2.1 B-spline for component functions

First we describe the spline approximation for component functions. Without loss of generality, we assume the support of x_{ij} is $[0,1]$ for $j = 1, \ldots, p$. Let $0 = \nu_0 < \nu_1 < \cdots < \nu_{K'-1} < \nu_{K'} = 1$ be a partition of $[0,1]$ into K' subintervals $I_{k'} = [\nu_{k'}, \nu_{k'+1}), k' = 0, \ldots, K'$, where K' increases with sample size n. In this paper, we simply take equally spaced knots (i.e., $\nu_k = k/K'$), while other data-driven choices of knots, for instance placing knots at sample quantiles of the observed values, may also be suitable.

Denote by S_n the space of polynomial splines of order $t \geq 2$. Any function $f(\cdot)$ from S_n satisfies: (i) on each $I_{k'}$, $1 \leq k' \leq K'$, $f(\cdot)$ is a polynomial of degree $t-1$; (ii) $f(\cdot)$ is globally $t-2$ times continuously differentiable on $[0,1]$. See the definition in Schumaker (2007) or Stone (1985). The collection of splines on $[0,1]$ with a fixed sequence of knots has a B-spline basis $\mathbf{B}(x) := \{B_1(x), \ldots, B_{K'+t}(x)\}$, which means that functions in S_n are of the form $\sum_{k=1}^{K'+t} a_k B_k(x)$ for some $a_k \in \mathbb{R}$. We assume the basis is scaled to have $\sum_{k=1}^{K'+t} B_k(x) = \sqrt{K'+t}$. Such normalization is not essential, but adopted to simplify some expressions in theoretical deductions later. In addition, we will then have the eigenvalues of $\int_0^1 \mathbf{B}(x)\mathbf{B}^{\mathrm{T}}(x)dx$ bounded away from zero and infinity, while if using the basis with $\sum_{k=1}^{K} B_k(x) = 1$ we would have the eigenvalues being of order $O((K')^{-1})$. Finally, to take into account the identifiability constraint $\int f_{jl}(x)dx = 0$, we center the basis functions to be $\mathbf{b}(x) = (b_1(x), \ldots, b_K(x))^{\mathrm{T}}$ with $K = K' + t - 1$ and $b_k(x) = B_k(x) - \int_0^1 B_k(x)dx$. Due to centering, the basis functions become linearly dependent and thus one of them, for specificity the last one, is removed.

Recall that in the multivariate model (1), the aim is to estimate the unknown functions $f_{jl}, j \in \{1, \ldots, p\}, l \in \{1, \ldots, q\}$. Using splines, we approximate $f_{jl}(x) \approx \mathbf{b}^{\mathrm{T}}(x)\mathbf{d}_{jl}$ where $\mathbf{d}_{jl} = (d_{j1l}, \ldots, d_{jKl})^{\mathrm{T}}$. When K increases with the sample size n at a reasonable order, the approximation error can be well controlled. For example, if f_{jl} is d times differentiable, it is known that there exist coefficients d_{jkl} such that $|f_{jl}(x) - \mathbf{b}^{\mathrm{T}}(x)\mathbf{d}_{jl}| = O(K^{-d})$ (Schumaker, 2007). In other words, as $K \to \infty$, the approximation can converge to zero, although in practice, it is rare to use K beyond 10 (Huang et al., 2010; Fan et al., 2014). Stone (1985) studied spline estimator for additive models with a fixed p and showed that the spline estimator can achieve the optimal convergence rate for functions in the Hölder class. However, when p and/or q are large, there are a large number of nonparametric functions to estimate, which makes statistical estimation inefficient unless the sample size is very large.

2.2 Spline coefficients as a low-rank tensor

We propose a natural and general framework for dimension reduction in multivariate additive models, by treating $\mathbf{D} = \{d_{jkl}\}$ as a third-order tensor and using an appropriate definition of low-rank tensors to achieve dimension reduction in the estimation.

Among many possible tensor decompositions proposed in the literature, CP decomposition and Tucker decomposition are the two most popular ones, both of which can be considered to be higher-order generalizations of the matrix singular value decomposition. Although in principle either of them can potentially be used to reduce the dimension of $\mathbf{D}$, we focus on Tucker decomposition here since it is more closely related to singular value

decomposition for matrices and thus more easily interpretable (de Lathauwer et al., 2000). Furthermore, several existing dimension reduction approaches in the recent literature can be regarded as special cases of using Tucker decomposition, as we will explain later in Section 2.3. However, none of these studies take advantage of the natural tensor structure of $\mathbf{D}$ but instead treat it as a low-rank matrix after a certain matricization operation.

The Tucker decomposition for the third order tensor $\mathbf{D}$ can be written as

$$\mathbf{D} = \mathbf{S} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}. \tag{3}$$

Without further constraints, the above decomposition can be trivial in that we can always choose $\mathbf{S} = \mathbf{D}$ and take $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ to be identity matrices. Such a decomposition is obviously useless. Under the low-rank assumption, we assume that the size of $\mathbf{S}$ is smaller than that of $\mathbf{D}$. More specifically, the low-rank assumption imposes that $\mathbf{D}$ of size $p \times K \times q$ can be decomposed into a *core tensor* $\mathbf{S}$ of size $r_1 \times r_2 \times r_3$ multiplied by a matrix along each mode, with $\mathbf{A}, \mathbf{B}, \mathbf{C}$ of sizes $p \times r_1$, $K \times r_2$ and $q \times r_3$, respectively. Such a low-rank assumption on tensors can be regarded as an extension of the low-rank assumption for matrices used in reduced-rank regression, in which a $p \times q$ matrix is assumed to be decomposable as the product of a $p \times r$ and a $r \times q$ matrix.

Note that the Tucker decomposition is not unique since for any non-singular square matrices $\mathbf{U}, \mathbf{V}, \mathbf{W}$ of suitable dimensions, we have

$$\mathbf{S} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = (\mathbf{S} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}) \otimes_1 (\mathbf{A}\mathbf{U}^{-1}) \times_2 (\mathbf{B}\mathbf{V}^{-1}) \times_3 (\mathbf{C}\mathbf{W}^{-1}). \tag{4}$$

This decomposition also means that we can, when necessary, assume that $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are orthogonal matrices, which will be convenient for some of our analysis later. Informally, the number of parameters to be estimated (degrees of freedom) in (3) is $pr_1 + Kr_2 + qr_3 + r_1r_2r_3 - r_1^2 - r_2^2 - r_3^2$, which can be much smaller than the number of elements pKq in $\mathbf{D}$. To see this, for simplicity we first assume the upper $r_1 \times r_1$ block of $\mathbf{A}$ is nonsingular and then by (4), we can assume without loss of generality that $\mathbf{A}$ takes the form $\mathbf{A} = [\mathbf{I}_{r_1 \times r_1}\ \mathbf{A}_1]^{\mathrm{T}}$. Thus in the estimation of $\mathbf{A}$, we only need to estimate the $pr_1 - r_1^2$ parameters in $\mathbf{A}_1$. The number of free parameters in $\mathbf{A}$ is thus $pr_1 - r_1^2$. Similarly, the number of free parameters for $\mathbf{B}$ and $\mathbf{C}$ are $Kr_2 - r_2^2$ and $qr_3 - r_3^2$, respectively. Thus the total number of free parameters, that is, the number of parameters to be estimated in $\mathbf{S} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$, is $pr_1 + Kr_2 + qr_3 + r_1r_2r_3 - r_1^2 - r_2^2 - r_3^2$, where $r_1r_2r_3$ is the number of parameters in $\mathbf{S}$. This value of degrees of freedom is also more rigorously established in terms of the entropy number in the asymptotic analysis below.

Let $\mathcal{D}(r_1, r_2, r_3)$ be the set of tensors $\mathbf{D}$ that can be decomposed as in (3). Using spline approximation, $\sum_j f_{jl}(x_{ij}) \approx \sum_{j=1}^{p} \sum_{k=1}^{K} b_k(x_{ij}) d_{jkl} = (\mathbf{D}_{(3)})_{l\cdot} \mathbf{z}_i$, where $\mathbf{D}_{(3)}$ is the matricization of $\mathbf{D}$ along its 3rd mode, $(\mathbf{D}_{(3)})_{l\cdot}$ is the lth row of $\mathbf{D}_{(3)}$, and $\mathbf{z}_i$ is a pK-vector with components $b_k(x_{ij}), 1 \le j \le p, 1 \le k \le K$. The least squares optimization problem for $\mathbf{D}$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^{\mathrm{T}}$ can thus be written as

$$\begin{aligned} & \min_{\boldsymbol{\mu}, \mathbf{D} \in \mathcal{D}(r_1, r_2, r_3)} \sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\mu} - \mathbf{D}_{(3)} \mathbf{z}_i\|^2 \\ = & \min_{\boldsymbol{\mu}, \mathbf{S}, \mathbf{A}, \mathbf{B}, \mathbf{C}} \sum_{i=1}^{n} \|\mathbf{y}_i - \boldsymbol{\mu} - (\mathbf{S} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C})_{(3)} \mathbf{z}_i)\|^2. \end{aligned} \tag{5}$$

Let $\widehat{\mathbf{D}}$ be the minimizer of (5). Since we are only interested in $\mathbf{D}$ rather than $\mathbf{S}, \mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, we do not need to be concerned about the identifiability issue and computationally we do not impose the constraint that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are orthogonal matrices. The strategy is to alternatingly update one of $\mathbf{S}, \mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ while fixing the others. Of course one can certainly use (4) to orthogonalize the $\mathbf{A}, \mathbf{B}$ or $\mathbf{C}$ after updating it to force it into this standardized form, if one chooses to.

2.3 Relation to some previous works

Our work can be regarded as a novel combination of tensor decomposition/optimization with semiparametric statistical modelling. There are certainly a large number of works on tensor decomposition/optimization, for example Anandkumar et al. (2012, 2014b); Sedghi and Sabharwal (2018). In the paper Hao et al. (2021), the authors studied sparse tensor additive regression. Besides that we have multivariate responses, the key difference is that the predictor considered in Hao et al. (2021) is tensorial and thus it is natural to use tensor method. However, in our case, we have a vector predictor $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and thus the method of Hao et al. (2021) does not apply to our case. It is traditionally not realized that the additive models can be approached using tensors, after series expansion. Thus our work represents a methodological advancement in using *tensor method for non-tensor models*. There are certainly other works on applying tensor methods to non-tensor models, including generalized linear models and some latent variable models (Anandkumar et al., 2012, 2014a; Sedghi et al., 2016; Janzamin et al., 2019).

Traditionally, estimation of high-dimensional additive models are focused on variable selection under the sparsity assumption (Ravikumar et al., 2008; Meier et al., 2009; Huang et al., 2010). These works do not use low-rank factorization and roughly correspond to our approach when $r_1 = p, r_2 = K, r_3 = q$. More related to our work, low-rankness assumptions were also adopted for semiparametric statistical models in a recently published work (He et al., 2022), which we became aware of after we received the first review of the current work. In He et al. (2022), the authors focused on univariate response additive models, and after using splines expansion for the unknown component functions, they assume the spline coefficient matrix is low-rank and matrix decomposition is used to reduce the number of parameters. Their method can only deal with one response variable and thus when there are possible relationships among multivariate responses, the method fails to incorporate such information. Besides, they did not provide statistical convergence rate in their paper, while our theory clearly shows the role of the reduced number of parameters in the convergence rate. In our numerical studies, we will also compare our tensor-based method to the method of He et al. (2022) to demonstrate the improved performance of the tensor-based method. Other works combining low-rank assumption with semiparametric modelling include Jiang et al. (2013); He et al. (2018) which also used matrix decomposition.

2.4 Asymptotic properties

In this section, we consider the asymptotic properties of the proposed estimators. The main goal of the analysis is to reveal the improved convergence rate due to the reduced effective number of parameters in the low-rank model. We make the following assumptions.

(A1) The joint density of $\mathbf{x}_i$ is bounded away from zero and infinity on $[0, 1]^p$.

(A2) The eigenvalues of $\sum_{i=1}^{n}(1,\mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}}(1,\mathbf{z}_i^{\mathrm{T}})/n$ are bounded away from zero and infinity with probability tending to one as $n \to \infty$.

(A3) The noise vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1},\ldots,\epsilon_{iq})^{\mathrm{T}}$ is independent of the predictor and is sub-Gaussian in the sense that $E[e^{a\boldsymbol{\epsilon}_i^{\mathrm{T}}\boldsymbol{\eta}}] \le e^{Ca^2\|\boldsymbol{\eta}\|^2}$ for any non-random $\boldsymbol{\eta} \in \mathbb{R}^q$.

(A4) For $g = f_{jl}, 1 \le j \le p, 1 \le l \le q$, g satisfies a Lipschitz condition of order $d > 1/2$: $|g^{(\lfloor d \rfloor)}(t) - g^{(\lfloor d \rfloor)}(s)| \le C|s-t|^{d-\lfloor d \rfloor}$, where $\lfloor d \rfloor$ is the largest integer strictly smaller than d and $g^{(\lfloor d \rfloor)}$ is the $\lfloor d \rfloor$-th derivative of g. The order of the B-spline used satisfies $t \ge d + 1/2$.

(A5) Let $\mathbf{D}_0 = (d_{0jkl})$ be the spline coefficients satisfying $|\sum_{k=1}^{K} b_k(x)d_{0jkl} - f_{jl}(x)| \le CK^{-d}$. We assume $\mathbf{D}_0$ satisfies (3), that is, $\mathbf{D}_0 \in \mathcal{D}(r_1, r_2, r_3)$.

These assumptions are generally mild. (A1) roughly means we will have a sufficient number of observations in all locations. If there is a low probability of predictors being observed in a certain location, then it is impossible to estimate the regression function near that location. The region $[0,1]^p$ can be changed to any rectangular region $\prod_{j=1}^{p}[a_j, b_j]$ by a simple scaling. Note that for spline-based estimation, it is common to assume the covariates are bounded in order to construct spline basis functions. Assumption (A2) on the eigenvalues of $\sum_{i=1}^{n}(1,\mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}}(1,\mathbf{z}_i^{\mathrm{T}})/n$ are frequently assumed. This is the same as that assumed in equations (35) and (36) of Ravikumar et al. (2008) in high-dimensional additive models. This is rigorously established as Lemma A.1 of Lian and Liang (2013) which showed that (A2) holds when $Kp\log(K \vee p)/n \to 0$. In the linear regression models, one typically assumes (or proves from other primitive assumptions) that the eigenvalues of $\sum_{i=1}^{n}(1,\mathbf{x}_i^{\mathrm{T}})^{\mathrm{T}}(1,\mathbf{x}_i^{\mathrm{T}})/n$ are bounded away from zero and infinity, which restricts the correlations among different predictors, and our assumption is simply the nonparametric counterpart of such an assumption. In (A3), sub-gaussianity of errors is assumed in order to apply Dudley's bound for the error term in our proof, which includes Gaussian error distribution as a special case. Sub-gaussianity is often assumed for theoretical studies of high-dimensional models (Wainwright, 2019). Assumption (A3) means our procedure does not work well for heavy-tailed error distribution or for data containing outliers, which is as expected since it is well-known that the least squares procedure is not robust. Robust version of the proposed approach should be used to deal with heavy-tailed errors, which is however not the main focus of the current work. Smoothness assumption in (A4) on the component functions is imposed to obtain nontrivial approximation rate for B-splines. This is a standard assumption even for $p = q = 1$ (nonparametric regression). Smoothness of the regression function is reasonably expected for many applications where the effects of the predictors change slowly with the value of the predictors. Assumption (A5) is fundamental to our analysis, which has been discussed in section 2.3. This is a direct extension of the low-rank assumption almost always used in matrix estimation problems. The low-rank assumption reduces the number of parameters. Bear in mind that without using the low-rank assumption, the number of parameters in $\mathbf{D}$ is pqK, estimation of which requires possibly a very large sample size. Such curse of dimensionality is lessened by imposing a low-rank approximation to the high-dimensional coefficient array. In real analysis, the predictors and the coefficient are probably not exactly low-rank. However, with a limited sample size, a low-rank estimate can provide a reasonable approximation to the true parameter, even when the truth is not

low-ranked. This phenomenon has been observed in some previous studies (Zhou et al., 2013) and also appears to be true in our real data illustration which showed the lower prediction error of the low-rank model. With minor modifications of the proof, our theoretical results also apply to the approximately low-rank case (see our Remark 2).

Theorem 1 *Under assumptions (A1)-(A5) and $K \to \infty$, we have*

$$\|\widehat{\mathbf{D}} - \mathbf{D}_0\| = O_p(\sqrt{df/n} + pqK^{-d}),$$

where $df = r_1r_2r_3 + pr_1 + Kr_2 + qr_3 - r_1^2 - r_2^2 - r_3^2$ can be deemed as the effective degrees of freedom under the low-rank assumption. In particular, this implies $\sum_{l,j} \|\widehat{f}_{jl} - f_{jl}\|^2 = O_p(df/n + p^2q^2K^{-2d})$.

Remark 1 *This result establishes the convergence rate of the difference between the coefficients in the spline representation of the nonparametric functions and their estimators. As a consequence, the rate of convergence of the estimated functions follows based on the properties of spline approximation, and we also have $\sum_{l,j} \|\widehat{f}_{jl} - f_{jl}\|^2 = O_p(df/n + p^2q^2K^{-2d})$. The two terms in the convergence rate correspond to the stochastic error and the (squared) approximation error, respectively. The theory shows there is a trade-off in the choice of the value of K. When K increases, the approximation error term pqK^{-d} decreases while the stochastic error increases. If we choose K to satisfy $K^{2d+1} \asymp \frac{np^2q^2}{r_2}$, then $\sum_{l,j} \|\widehat{f}_{jl} - f_{jl}\|^2 = O_p(df/n)$. When $p = q = 1$ and r_2 is fixed, this choice of K leads to the convergence rate $n^{-\frac{2d}{2d+1}}$ which is the minimax convergence rate for nonparametric regression (Tsybakov, 2009).*

Remark 2 *Our result can be straightforwardly extended to the case that the true coefficient tensor is approximately low-rank in the sense that $\min_{\mathbf{D}\in\mathcal{D}(r_1,r_2,r_3)} \|\mathbf{D}_0 - \mathbf{D}\| \le \xi$ for some $\xi > 0$. With other assumptions unchanged, we can show the convergence rate is $\|\widehat{\mathbf{D}} - \mathbf{D}_0\| = O_p(\sqrt{df/n} + pqK^{-d} + \xi)$. The proof of this bound is also contained in the proof of Theorem 1 in the appendix.*

3 Tensor decomposition combined with sparse modelling

The low-rank tensor decomposition method proposed in the previous section reduces the number of parameters from pKq to $r_1r_2r_3 + pr_1 + Kr_2 + qr_3 - r_1^2 - r_2^2 - r_3^2$. However, for example, the number of parameters still increases linearly with the covariate dimension p. When the covariate dimension is high, a standard method that has proved powerful is to incorporate variable selection methodology under a sparsity assumption. In this section we consider penalized estimation for variable selection. We assume without loss of generality that only the first s predictors are useful. In other words, we assume $f_{jl} \equiv 0$ for $j > s$. One can also consider the possibility that each response variable may have a different set of useful predictors, which can be implemented using a slightly different way of penalizing parameters than used here. But this alternative is not further considered in this work.

In the coefficient tensor $\mathbf{D}$, the slice (a $K \times q$ matrix) $\mathbf{D}_{j\cdot\cdot}$ is the part that is associated with predictor j and thus one strategy for variable selection is to shrink $\|\mathbf{D}_{j\cdot\cdot}\|$ to zero. Let

$\mathbf{d}_j$ be the vectorization of $\mathbf{D}_{j..}$. We can certainly shrink $\|\mathbf{F}_j\mathbf{d}_j\|$ to zero for any $\mathbf{F}_j$ without changing much of the theoretical aspect of the estimator as long as the maximum and the minimum eigenvalues of $\mathbf{F}_j^{\mathrm{T}}\mathbf{F}_j$ are of the same order. If the maximum and the minimum eigenvalues of $\mathbf{F}_j^{\mathrm{T}}\mathbf{F}_j$ are not of the same order, the proof can also be modified. However, in this case, $\|\mathbf{F}_j\mathbf{d}_j\|$ will penalize certain vector directions much more heavily than others and thus it seems harder to justify the use of such $\mathbf{F}_j$. In this work, we will penalize $\|\mathbf{Z}_j\mathbf{d}_j/\sqrt{n}\|$ instead of $\|\mathbf{d}_j\|$, where $\mathbf{Z}_j$ contains the columns of $\mathbf{Z}$ associated with the j-th predictor with $\mathbf{Z}_{nq\times pKq} = (\mathbf{z}_1 \otimes \mathbf{I}_q, \ldots, \mathbf{z}_n \otimes \mathbf{I}_q)^{\mathrm{T}}$, resulting in the optimization problem

$$\min_{\mathbf{D}\in\mathcal{D}(r_1,r_2,r_3)} \|\bar{\mathbf{y}} - \mathbf{1}\otimes\boldsymbol{\mu} - \mathbf{Z}\mathbf{d}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}), \tag{6}$$

where $\bar{\mathbf{y}} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_n^{\mathrm{T}})^{\mathrm{T}}$ is nq-vector of responses, $\mathbf{1}$ is a n-vector of ones, $\mathbf{d} = \mathrm{vec}(\mathbf{D}_{(3)})$. Here $p_\lambda(.)$ is the penalty function and λ is the associated tuning parameter that controls the amount of shrinkage in parameters. For example, if we use $p_\lambda(x) = \lambda|x|$, this is just the ℓ_1/lasso penalty (Tibshirani, 1996).

For the high-dimensional models, variable selection plays a vital role in identifying the relevant and useful features and improving the accuracy of estimation. Many variable selection techniques have been developed, such as bridge regression (Frank and Friedman (1993)), the lasso (Tibshirani (1996)), the adaptive lasso (Zou (2006)) and the Dantzig selector (Candes and Tao (2007)). There are also nonconvex penalties proposed, including the smoothly clipped absolute deviation (SCAD, Fan and Li (2001)), and the minimax concave penalty (MCP, Zhang (2010)). It is known that the lasso penalty generally do not have variable selection consistency (that is, it does not correctly separately the nonzero and zero parameters as $n \to \infty$), while the SCAD/MCP penalty can have such properties. Thus, in this paper, we mainly consider these two penalties theoretically, but we also use the ℓ_1 penalty in our numerical studies. The SCAD penalty is defined as

$$p_\lambda(|u|) = \lambda|u|I(0 \le |u| < \lambda) + \frac{a\lambda|u| - (|u|^2+\lambda^2)/2}{a-1} I(\lambda \le |u| \le a\lambda) + \frac{(a+1)\lambda^2}{2} I(|u| > a\lambda),$$

for some $a > 2$, and the MCP defined as

$$p_\lambda(|u|) = \left(\lambda|u| - \frac{u^2}{2a}\right) I(|u| \le a\lambda) + \frac{a\lambda^2}{2} I(|u| > a\lambda),$$

for some $a > 1$.

The specific choice of penalizing $\|\mathbf{Z}_j\mathbf{d}_j\|$ instead of $\|\mathbf{d}_j\|$ is motivated from two aspects. First, $\mathbf{Z}_j\mathbf{d}_j$ is the component of the conditional mean that represents directly the effect of predictor j on the responses. Second, it allows an efficient group descent algorithm proposed in Breheny and Huang (2015) as will be made clear below.

Theorem 2 *Under assumptions (A1)-(A5), with* $\mathbf{z}_i$ *in (A2) understood as the vector* $b_k(x_{ij})$ *based on only* $j \le s$*, and that*

$$\frac{\log(pKq)}{n} + \frac{df}{n} + s^2q^2K^{-2d} \ll \lambda^2 \ll \min_{j\le s}\sum_{l=1}^{q}\|f_{jl}\|^2,$$

where we used the Vinogradov symbol $\ll$ such that $a \ll b$ means $a = o(b)$, there is a local minimizer of (6) $\breve{\mathbf{D}}$*, using either SCAD or MCP as the penalty, such that*

$$\|\breve{\mathbf{D}} - \mathbf{D}_0\| = O_p(\sqrt{df/n} + sqK^{-d}),$$

and $\mathbf{D}_{j\cdot\cdot}$ is a zero matrix for $j > s$ with probability approaching one, where $df = r_1r_2r_3 + r_1(s-r_1)+r_2(K-r_2)+r_3(q-r_3)$. In particular, this implies $\sum_{l,j}\|\breve{f}_{jl} - f_{jl}\|^2 = O_p(df/n + s^2q^2K^{-2d})$ and $\breve{f}_{jl} = 0$ for all $j > s$ with probability approaching one.

Remark 3 *This theorem shows the existence of a local minimizer of the objective function with the oracle property, in the sense that it correctly selects the nonzero functions and achieves the optimal rate of convergence. In particular, if K satisfies $K^{2d+1} = \frac{ns^2q^2}{r_2}$, then $\sum_{l,j}\|\breve{f}_{jl} - f_{jl}\|^2 = O_p(df/n)$.*

4 Computation

We will use Einstein summation notation. That is, repeated indices (one upper and one lower) are summed over. Denote $b_i^{jk} = b_k(x_{ij})$ and denote entries of $\mathbf{A}, \mathbf{B}, \mathbf{C}$ by $\mathbf{A}_j^a, \mathbf{B}_k^b, \mathbf{C}_l^c$, respectively. Then the minimization problem is

$$\sum_{i,l}(y_{il} - \mathbf{s}_{abc}\mathbf{A}_j^a\mathbf{B}_k^b\mathbf{C}_l^c b_i^{jk})^2$$

It is clear that $\mathbf{s}_{abc}\mathbf{A}_j^a\mathbf{B}_k^b\mathbf{C}_l^c b_i^{jk}$ is linear in one of $\mathbf{S}, \mathbf{A}, \mathbf{B}, \mathbf{C}$ when the other three are fixed. Thus alternating updates can be easily implemented.

For the penalized estimators, we also use the alternating updating strategy, but due to the presence of the penalty function, the computation is more complicated. In our numerical studies, we will consider SCAD and MCP penalties, as well as the group lasso penalty. For illustration, consider only the updating of $\mathbf{A}$. Since $\mathbf{D}_{(1)} = \mathbf{A}\mathbf{S}_{(1)}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}}$, we have $\mathbf{d}_j = (\mathbf{C}\otimes\mathbf{B})\mathbf{S}_{(1)}^{\mathrm{T}}\mathbf{a}_j$, where $\mathbf{a}_j$ is the j-th row of $\mathbf{A}$ (as a column vector). Thus to update $\mathbf{a}_j$ we write (6) as

$$\sum_{i=1}^n\|\bar{\mathbf{y}} - \sum_{j'\neq j}\mathbf{Z}_{j'}\mathbf{d}_{j'} - \mathbf{Z}_j(\mathbf{C}\otimes\mathbf{B})\mathbf{S}_{(1)}^{\mathrm{T}}\mathbf{a}_j\|^2 + n\sum_{j=1}^p p_\lambda(\|\mathbf{Z}_j(\mathbf{C}\otimes\mathbf{B})\mathbf{S}_{(1)}^{\mathrm{T}}\mathbf{a}_j\|/\sqrt{n}).$$

Let $\mathbf{\Gamma}_j = \mathbf{S}_{(1)}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j(\mathbf{C}\otimes\mathbf{B})\mathbf{S}_{(1)}^{\mathrm{T}}/n$, $\mathbf{a}_j^{new} = \mathbf{\Gamma}_j^{1/2}\mathbf{a}_j$, $\mathbf{Z}_j^{new} = \mathbf{Z}_j(\mathbf{C}\otimes\mathbf{B})\mathbf{S}_{(1)}^{\mathrm{T}}\mathbf{\Gamma}_j^{-1/2}$, the above is equivalent to

$$\sum_{i=1}^n\|\bar{\mathbf{y}} - \sum_{j'\neq j}\mathbf{Z}_{j'}\mathbf{d}_{j'} - \mathbf{Z}_j^{new}\mathbf{a}_j^{new}\|^2 + n\sum_{j=1}^p p_\lambda(\|\mathbf{a}_j^{new}\|),$$

with $(\mathbf{Z}_j^{new})^{\mathrm{T}}\mathbf{Z}_j^{new}/n = \mathbf{I}$. Then the group descent algorithm of Breheny and Huang (2015) can be used to update $\mathbf{a}_j, j = 1, \ldots, p$.

More specifically, we can implement the updating of $\mathbf{A}$ as follows. Let $\mathrm{tr}_j = \bar{\mathbf{y}} - \sum_{j' \neq j} \mathbf{Z}_{j'} \mathbf{d}_{j'}$ and write the above as

$$Q(\mathbf{a}_j^{new}) = \left\| \mathrm{tr}_j - \mathbf{Z}_j^{new} \mathbf{a}_j^{new} \right\|^2 /2 + n \sum_{j=1}^{p} p_\lambda(\|\mathbf{a}_j^{new}\|).$$

If the group lasso penalty is used, we have the first-order conditions

$$\frac{\partial Q(\mathbf{a}_j^{new})}{\partial \mathbf{a}_j^{new}} = \begin{cases} -(\mathbf{Z}_j^{new})^{\mathrm{T}} \mathrm{tr}_j + n\mathbf{a}_j^{new} + n\frac{\lambda}{\|\mathbf{a}_j^{new}\|} \mathbf{a}_j^{new}, & \text{if } \mathbf{a}_j \neq \mathbf{0}; \\ -(\mathbf{Z}_j^{new})^{\mathrm{T}} \mathrm{tr}_j + n\lambda \mathbf{v}, & \text{if } \mathbf{a}_j^{new} = \mathbf{0}, \end{cases}$$

where $\mathbf{v}$ is any r_1-vector satisfying $\|\mathbf{v}\| \leq 1$. Let $\tilde{\boldsymbol{z}}_j = n^{-1}(\mathbf{Z}_j^{new})^{\mathrm{T}} \mathrm{tr}_j$. The solution is easily derived to be

$$\hat{\mathbf{a}}_j^{new} = S(\tilde{\boldsymbol{z}}_j, \lambda) \frac{\tilde{\boldsymbol{z}}_j}{\|\tilde{\boldsymbol{z}}_j\|},$$

where $S(z, \lambda) = (|z| - \lambda)_+$ is the soft thresholding operator.

The MCP and SCAD estimators can be calculate similarly. Define

$$f_{\mathrm{mcp}}(z, \lambda, a) = \begin{cases} \frac{S(z,\lambda)}{1-1/a}, & \text{if } |z| \leq a\lambda; \\ z, & \text{if } |z| > a\lambda. \end{cases}$$

and

$$f_{\mathrm{scad}}(z, \lambda, a) = \begin{cases} S(z, \lambda), & \text{if } |z| \leq 2\lambda; \\ \frac{S(z, a\lambda/(a-1))}{1-1/(a-1)}, & \text{if } 2\lambda < |z| \leq a\lambda; \\ z, & \text{if } |z| > a\lambda. \end{cases}$$

Replacing $S(z, \lambda)$ by $f_{\mathrm{mcp}}(z, \lambda, a)$ and $f_{\mathrm{scad}}(z, \lambda, a)$ yields the algorithm for updating of $\mathbf{A}$ with MCP and SCAD penalties, respectively. More details can be found in Breheny and Huang (2015). We have implemented the algorithm described above in an R package `tensorMAM` available at (`https://github.com/xliusufe/tensorMam`).

5 Numerical studies

5.1 Simulations

In this section we perform simulation studies to illustrate the performance of the proposed estimator. The predictors x_{ij} are generated independently from the uniform distribution $U(0, 1)$. For the tensor, we set $(r_1, r_2, r_3) = (2, 2, 2)$, each entry of $\mathbf{S}$ is generated independently from $U(10, 20)$ if $s = 40$ and from $U(3, 7)$ if $s = 20$ (s is the number of significant predictors). Orthogonal matrices $\mathbf{U}_j$ are uniformly drawn from the Stiefel manifold (the set of orthogonal matrices of a given size), which can be achieved by first generating a matrix $\mathbf{T}_j$ with entries i.i.d. from $N(0, 1)$ and then perform a QR decomposition to get the orthogonal matrix. We consider two scenarios for the response:

Sce 1 We set $\mathbf{D}_{s\times K\times q} = \mathbf{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$. Given the tensor $\mathbf{D}$, we set the functions $f_{jl}(x_{ij}) = \mathbf{d}_{jl}^{\mathrm{T}}\mathbf{b}(x_{ij})$ and add i.i.d. errors generated from $N(0, \sigma^2)$. In this setting, the functions are actually in the space of splines.

Sce 2 Define two functions $g_1(x) = \sin(2\pi x)$ and $g_2(x) = \cos(\pi x)$ and let $\mathbf{g}_1 = (g_1(x_{ij}))$ and $\mathbf{g}_2 = (g_2(x_{ij}))$ be $n \times s$ matrices. Let $\mathbf{S}_1, \mathbf{S}_2$ be the two $s \times q$ slices of $\mathbf{S} \times_1 \mathbf{U}_1 \times_3 \mathbf{U}_3$. Finally, we set

$$\mathbf{y}_{n\times q} = \mathbf{g}_1 \times \mathbf{S}_1 + \mathbf{g}_2 \times \mathbf{S}_2 + \mathbf{E}_{n\times q},$$

where the errors are generated as in Scenario 1.

For each scenario, 500 datasets are generated with the number of responses $q = 10$. We set $n = 200, 500$ and $p = 100$ (only the first s predictors are useful), and $\sigma^2 = 0.2$. We use cubic B-splines with 3 internal knots. Although one might try to select the number of knots based on data, using fixed knots is convenient in practice and often adopted in the literature. We select the ranks as well as λ when fitting the proposed models using BIC, where the degrees of freedom in BIC is as defined in the statement of Theorem 2. For illustration, we consider three penalties including (group) lasso, SCAD and MCP.

The results for the estimated ranks $(\hat{r}_1, \hat{r}_2, \hat{r}_3)$ are reported in Table 1. We see that when $n = 200$, the method tends to under-estimate the ranks. However, this improves with larger sample size $n = 500$. This behavior is reasonable since with a smaller sample size a more parsimonious model with fewer parameters is preferred.

To examine the performance of variable selection, we also compare our method with the oracle estimator and the full-rank (FR) estimator. Here the oracle estimator is the one which fixes the ranks to be $(2, 2, 2)$ while still using a penalty to select significant variables, and the full-rank estimator does not use low-rank factorization. We also compare our estimator with that of the high-dimensional additive models (HAM) (Meier et al., 2009), the sparse additive models (SAM) (Ravikumar et al., 2009), and the reduced additive models with the group Lasso (RAM_GL) and with the adaptive group Lasso (RAM_AGL) (He et al., 2022). Table 2 reports the True Positives (TP, the number of correctly estimated nonzero coefficients), the True Negatives (TN, the number of correctly estimated zero coefficients), the False Positives (FP, the number of incorrectly estimated nonzero coefficients), the True Negatives (TN, the number of correctly estimated zero coefficients) and the False Negatives (FN, the number of incorrectly estimated zero coefficients), Model Sizes (MS, the number of selected variables), and Matthews Correlation Coefficient (MCC), defined as

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

A larger MCC value indicates better variable selection performance. From Table 2, we can see that the results for variable selection are generally satisfactory with high MCC values. Compared with MCP and SCAD estimators, Lasso estimators tend to have slightly larger model sizes with larger false positives. FR, HAM, SAM, RAM_GL and RAM_AGL all produce much smaller MCC compared to the proposed method.

Finally, we calculate the integrated mean squared errors (IMSE) for the estimated function of $f_{jl}(\cdot)$, defined as $\mathrm{IMSE} = \int_0^1 (\widehat{f}_{jl}(x) - f_{jl}(x))^2 dx$. Table 3 reports the median of the square root of IMSEs for the proposed estimator, FR, HAM, SAM, RAM_GL and

RAM_AGL. The estimators based on the MCP and the SCAD penalties have smaller IMSE compared to the lasso estimator. Furthermore, the proposed estimator performs much better than both the full-rank estimator which does not perform dimension reduction and the reduced additive models with the group Lasso (RAM_GL) and with the adaptive group Lasso (RAM_AGL) which do not reduce ranks as an order-three tensor.

Table 1: The frequency of the estimated ranks $(\hat{r}_1, \hat{r}_2, \hat{r}_3)$ of tensor $\mathbf{D}$ in models with sparsity.

(n,s)	$\hat{r}_j$	LASSO					MCP					SCAD				
		=1	=2	=3	=4	=5	=1	=2	=3	=4	=5	=1	=2	=3	=4	=5
		Scenario 1														
	$\hat{r}_1$	500	0	0	0	0	500	0	0	0	0	500	0	0	0	0
(200,20)	$\hat{r}_2$	493	7	0	0	0	499	1	0	0	0	499	1	0	0	0
	$\hat{r}_3$	500	0	0	0	0	500	0	0	0	0	500	0	0	0	0
	$\hat{r}_1$	500	0	0	0	0	500	0	0	0	0	500	0	0	0	0
(200,40)	$\hat{r}_2$	457	38	3	2	0	494	5	1	0	0	492	7	0	1	0
	$\hat{r}_3$	491	9	0	0	0	500	0	0	0	0	500	0	0	0	0
	$\hat{r}_1$	92	408	0	0	0	23	477	0	0	0	29	471	0	0	0
(500,20)	$\hat{r}_2$	92	407	1	0	0	23	477	0	0	0	29	471	0	0	0
	$\hat{r}_3$	92	408	0	0	0	23	477	0	0	0	29	471	0	0	0
	$\hat{r}_1$	500	0	0	0	0	500	0	0	0	0	500	0	0	0	0
(500,40)	$\hat{r}_2$	95	403	2	0	0	57	443	0	0	0	63	437	0	0	0
	$\hat{r}_3$	95	405	0	0	0	57	443	0	0	0	63	437	0	0	0
		Scenario 2														
	$\hat{r}_1$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0
(200,20)	$\hat{r}_2$	0	493	7	0	0	0	500	0	0	0	0	500	0	0	0
	$\hat{r}_3$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0
	$\hat{r}_1$	500	0	0	0	0	494	6	0	0	0	500	0	0	0	0
(200,40)	$\hat{r}_2$	97	328	58	17	0	64	427	7	2	0	103	390	6	1	0
	$\hat{r}_3$	104	396	0	0	0	64	436	0	0	0	104	396	0	0	0
	$\hat{r}_1$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0
(500,20)	$\hat{r}_2$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0
	$\hat{r}_3$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0
	$\hat{r}_1$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0
(500,40)	$\hat{r}_2$	0	493	7	0	0	0	500	0	0	0	0	500	0	0	0
	$\hat{r}_3$	0	500	0	0	0	0	500	0	0	0	0	500	0	0	0

5.2 Breast cancer data example

We applied the proposed method to a breast cancer dataset (Chin et al. (2006), Witten et al. (2009) and Chen et al. (2013)), which includes gene expressions and comparative genomic hybridization measurements for 89 subjects. For illustration, we selected chromosome 21,

Table 2: The performances of variable selection in terms of TP, TN, FP, FN, MS, and MCC.

	(n,s)	Scenario 1						Scenario 2					
		TP	TN	FP	FN	MS	MCC	TP	TN	FP	FN	MS	MCC
						LASSO							
Oracle	(200,20)	17.25	77.55	2.45	2.75	19.70	0.84	18.36	78.81	1.19	1.64	19.55	0.91
	(200,40)	31.74	53.12	6.88	8.26	38.62	0.69	28.83	57.09	2.91	11.17	31.75	0.71
	(500,20)	17.96	79.70	0.30	2.04	18.26	0.93	19.03	79.58	0.42	0.97	19.45	0.96
	(500,40)	36.95	58.78	1.22	3.05	38.16	0.91	37.59	58.60	1.40	2.41	38.99	0.92
Proposed	(200,20)	16.70	78.51	1.49	3.30	18.19	0.85	18.36	79.16	0.84	1.64	19.19	0.92
	(200,40)	34.14	55.24	4.76	5.86	38.90	0.78	26.63	57.09	2.91	13.37	29.55	0.67
	(500,20)	17.82	79.71	0.29	2.18	18.10	0.92	19.02	79.51	0.49	0.98	19.51	0.95
	(500,40)	36.31	58.30	1.70	3.69	38.01	0.89	37.55	58.96	1.04	2.45	38.60	0.93
FR	(200,20)	5.86	79.94	0.06	14.14	5.92	0.49	7.54	79.99	0.01	12.46	7.55	0.57
	(200,40)	8.43	59.69	0.31	31.57	8.74	0.36	9.24	59.70	0.30	30.76	9.54	0.38
	(500,20)	5.87	80	0	14.13	5.87	0.50	8.02	80	0	11.98	8.02	0.59
	(500,40)	9.47	60	0	30.53	9.47	0.40	10.15	60	0	29.85	10.15	0.41
RAM_GL	(200,20)	18.98	42.29	37.71	1.02	56.69	0.39	14.64	79.53	0.47	5.36	15.11	0.81
	(200,40)	38.20	15.30	44.70	1.80	82.90	0.27	25.71	52.86	7.14	14.29	32.85	0.55
	(500,20)	19.49	44.20	35.80	0.51	55.29	0.42	16.10	80.00	0.00	3.90	16.10	0.88
	(500,40)	39.49	12.92	47.08	0.51	86.57	0.29	29.98	59.97	0.03	10.02	30.01	0.80
RAM_AGL	(200,20)	19.71	34.70	45.30	0.29	65.01	0.35	17.09	78.88	1.12	2.91	18.21	0.87
	(200,40)	39.28	13.13	46.87	0.72	86.15	0.28	23.76	57.64	2.36	16.24	26.12	0.62
	(500,20)	19.66	41.11	38.89	0.34	58.55	0.40	17.61	80.00	0.00	2.39	17.61	0.92
	(500,40)	39.75	12.33	47.67	0.25	87.42	0.29	28.19	60.00	0.00	11.81	28.19	0.77
SAM	(200,20)	8.43	78.58	1.42	11.57	9.85	0.54	13.22	77.67	2.33	6.78	15.55	0.70
	(200,40)	20.62	49.72	10.28	19.38	30.90	0.36	17.45	53.92	6.08	22.55	23.53	0.39
	(500,20)	5.43	80.00	0.00	14.57	5.43	0.48	9.22	80.00	0.00	10.78	9.22	0.64
	(500,40)	19.26	59.50	0.50	20.74	19.76	0.58	12.50	60.00	0.00	27.50	12.50	0.46
HAM	(200,20)	0.02	80.00	0.00	19.98	0.02	0.03	0.16	80.00	0.00	19.84	0.16	0.08
	(200,40)	6.38	59.90	0.10	33.62	6.48	0.31	0.13	60.00	0.00	39.87	0.13	0.04
	(500,20)	15.73	79.29	0.71	4.27	16.44	0.84	18.17	78.39	1.61	1.83	19.78	0.89
	(500,40)	6.22	59.99	0.01	33.78	6.23	0.31	0.04	60.00	0.00	39.96	0.04	0.02
						MCP							
Oracle	(200,20)	16.74	80	0	3.26	16.74	0.90	17.54	80	0	2.46	17.54	0.92
	(200,40)	29.68	59.91	0.09	10.32	29.77	0.79	26.96	59.82	0.18	13.04	27.14	0.74
	(500,20)	17	80	0	3	17	0.91	18.88	80	0	1.12	18.88	0.96
	(500,40)	34.43	60	0	5.57	34.43	0.89	36.67	59.98	0.02	3.33	36.69	0.93
Proposed	(200,20)	16.22	79.97	0.03	3.78	16.25	0.88	17.41	80	0	2.59	17.41	0.92
	(200,40)	31.79	59.98	0.02	8.21	31.81	0.84	24.73	59.80	0.20	15.27	24.94	0.70
	(500,20)	16.98	80	0	3.02	16.98	0.90	18.89	80	0	1.11	18.89	0.97
	(500,40)	34.38	60	0	5.62	34.38	0.89	36.56	60	0	3.44	36.56	0.93
FR	(200,20)	5.13	80	0	14.87	5.13	0.46	6.22	80	0	13.78	6.22	0.51
	(200,40)	6.80	59.93	0.07	33.20	6.87	0.33	7.40	59.93	0.07	32.60	7.47	0.34
	(500,20)	5.29	80	0	14.71	5.29	0.47	7.10	80	0	12.90	7.10	0.55
	(500,40)	8.26	60	0	31.74	8.26	0.37	8.95	60	0	31.05	8.95	0.38
						SCAD							
Oracle	(200,20)	17.17	79.86	0.14	2.83	17.30	0.91	18.41	79.87	0.13	1.59	18.54	0.95
	(200,40)	32.89	59.81	0.19	7.11	33.07	0.85	32.23	59.15	0.85	7.77	33.08	0.83
	(500,20)	17.31	80	0	2.69	17.31	0.92	18.98	80	0	1.02	18.98	0.97
	(500,40)	36.31	59.96	0.04	3.69	36.34	0.92	37.46	59.76	0.24	2.54	37.70	0.94
Proposed	(200,20)	16.71	79.76	0.24	3.29	16.95	0.89	18.31	79.98	0.02	1.69	18.33	0.95
	(200,40)	33.60	59.64	0.36	6.40	33.96	0.86	26.55	59.05	0.95	13.45	27.50	0.71
	(500,20)	17.21	80	0	2.79	17.21	0.91	18.99	80	0	1.01	18.99	0.97
	(500,40)	35.49	59.93	0.07	4.51	35.55	0.91	37.42	59.82	0.18	2.58	37.60	0.94
FR	(200,20)	5.85	79.94	0.06	14.15	5.90	0.49	7.54	79.99	0.01	12.46	7.55	0.57
	(200,40)	8.43	59.69	0.31	31.57	8.74	0.36	9.24	59.70	0.30	30.76	9.54	0.38
	(500,20)	5.82	80	0	14.18	5.82	0.50	8.01	80	0	11.99	8.01	0.59
	(500,40)	9.46	60	0	30.54	9.46	0.40	10.15	60	0	29.85	10.15	0.41

Table 3: The median of the square root of IMSE of the estimated functions.

Methods	Scenario 1 (n, s)				Scenario 2 (n, s)			
	(200,20)	(200,40)	(500,20)	(500,40)	(200,20)	(200,40)	(500,20)	(500,40)
	LASSO							
Proposed	7.61e-04	1.21e-03	2.72e-04	7.83e-04	2.88e-03	7.21e-04	2.12e-03	1.76e-03
FR	0.0754	0.124	0.0755	0.125	0.125	0.0874	0.119	0.0876
RAM_GL	0.1093	0.2211	0.1135	0.2221	0.0362	0.0464	0.0199	0.0253
RAM_AGL	0.1084	0.2215	0.1106	0.2211	0.0463	0.0626	0.0326	0.0492
SAM	0.1921	0.5576	0.1289	0.5283	0.4777	0.3044	0.4088	0.2789
HAM	0.0560	0.1807	0.0859	0.1795	0.1145	0.1148	0.0750	0.1149
	MCP							
Proposed	2.15e-04	7.86e-04	5.37e-05	1.92e-04	7.73e-04	5.93e-04	7.92e-04	5.63e-04
FR	0.0739	0.124	0.0697	0.122	0.115	0.0873	0.112	0.0856
	SCAD							
Proposed	1.94e-04	8.71e-04	1.18e-04	1.56e-04	6.12e-04	7.83e-04	8.64e-04	6.64e-04
FR	0.0754	0.124	0.0758	0.125	0.125	0.0874	0.119	0.0876

including $q = 44$ variables for copy-number variations and $p = 227$ variables for gene expression. As in Chen et al. (2013), we consider copy-number variations as the responses and gene expressions as the predictors. To illustrate the selection stability of the proposed method, we randomly split data into two parts multiple times, each with a training set containing 79 subjects and a test set containing 10 subjects. We standardize the responses to have mean zero and variance one, and transform the predictors into $[0, 1]$.

We use the proposed method with MCP penalty to select important genes for each random partition of the data. There are a total of 90 genes that are ever selected in 100 random splits. We list in Table 4 the top 6 genes with the highest frequency of being selected.

The top two genes have been shown to have significant associations between genome copy numbers measured using CGH and gene transcript level measured using Affymetrix U133A expression arrays in 101 primary breast tumors, of which the Pearson correlations are respectively 0.53 and 0.46, see Table S3 in the Supplementary Material of Chin et al. (2006). In Table 4, there are two segments of the same gene named "SON1" and "SON2" that have the same nucleotide position. The average prediction error (PE) is 1.90 and the average number of selected genes is 4.68.

We also compare the method with the linear reduced rank regression model of Chen et al. (2013) and the reduced additive model with the group Lasso (RAM_GL) and the adaptive group Lasso (RAM_AGL) (He et al. (2022)). As demonstrated in He et al. (2022) and the simulation studies in the previous section, HAM and SAM perform much worse than RAM_GL and RAM_AGL. Thus, we do not compute HAM and SAM here for the real data set. As in Chen et al. (2013), since the number of predictors is much larger than n, we first applied reduced rank regression with sparse singular value decomposition (Chen et al., 2012). Then, based on the selected predictors, the linear reduced rank regression method was applied using R package rrpack. The prediction error on the test data based on 100 random

splits is 5.96, which is larger than the prediction error of our proposed method. Notably, the average number of selected genes based on the linear reduced rank model is 46.73, which is about 10 times the number of the selected genes using our proposed method. Therefore, with a much smaller number of genes, our proposed method achieves a better prediction performance in comparison to the linear reduced rank regression model. Both RAM_GL and RAM_AGL were applied using R package fAdditiveModel. The prediction errors of RAM_GL and RAM_AGL based on 100 random splits are 1.33 and 1.13 respectively, but the average numbers of selected genes are 226.83 and 88.79, respectively, which are much greater than our proposed method.

Table 4: List of genes most frequently selected with their selected frequency among 100 random splits of the data.

gene name	"HRMT1L1"	"SON1"	"LSS"	"SON2"	"C21orf59"	"SMT3H1"
position	46911	33835	46465	33835	32894	45081
times selected	59	39	38	34	34	26

6 Conclusion

In this paper, we have proposed a tensor-based approach for multivariate additive models which involves a large number of component functions. Tensor decomposition allows a dramatic reduction of the number of parameters along all three directions of the tensor that results in improved efficiency of the estimation. We further combine dimension reduction with penalized variable selection to deal with ultra-high dimensionality. Our numerical studies have shown superior performance of the proposed estimators.

The current research can be extended in a few directions. First, the generalized linear model can deal with binary and count responses and represents another direction to make the linear model more flexible. Again, its application is partially limited by the linear form imposed. One can thus consider generalized additive models (Hastie and Tibshirani, 1990) with various response types, using the estimation strategy proposed here to make the estimation more accurate in some settings. One can also consider quantile additive models (Horowitz and Lee, 2005) to investigate the conditional distribution of the response at different quantile levels. Second, the current variable selection approach assumes a predictor is selected as long as it has an effect on one of the responses for the convenience of theory and implementation. In practice, it is more realistic to allow a different set of active predictors for different responses, possibly based on other types of penalty. These topics can be investigated in the future.

Acknowledgements

The authors sincerely thank the editor, associate editor, and two reviewers for their insightful and constructive comments that significantly improved the manuscript. Liu's research is supported by the NSFC (12271329, 72331005), the Program for Innovative Research Team of SUFE, the Shanghai Research Center for Data Science and Decision Technology, the Open Research Fund of Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, and the Open Research Fund of Key Laboratory of Analytical Mathematics and Applications (Fujian Normal University), Ministry of Education, P. R. China. Lian's research is supported by NSFC 12371297 at CityU Shenzhen Research Institute, NSF of Jiangxi Province under Grant 20223BCJ25017, and by Hong Kong RGC general research fund 11300519 and 11311822, and by CityU internal grant 7006014. The work of J. Huang is supported by the National Natural Science Foundation of China grant (No. 72331005) and research grants from The Hong Kong Polytechnic University.

Appendix: Proofs

In this appendix, we prove Theorems 1 and 2.

Proof of Theorem 1. For simplicity of notations, we first assume the intercepts are non-existent and at the end of the proof mention why this can be assumed without loss of generality. By the definition of the estimator, we have

$$\sum_i \|\mathbf{y}_i - \widehat{\mathbf{D}}_{(3)}\mathbf{z}_i\|^2 \le \sum_i \|\mathbf{y}_i - \mathbf{D}_{0(3)}\mathbf{z}_i\|^2.$$

Define $\mathbf{r}_i = (r_{i1}, \ldots, r_{iq})^{\mathrm{T}}$ where $r_{il} = \sum_j f_{jl}(x_{ij}) - \sum_{j,k} d_{0jkl} b_k(x_{ij})$ are the spline approximation errors. Rewriting the above using that $\mathbf{y}_i = \boldsymbol{\epsilon}_i + \mathbf{r}_i + (\mathbf{D}_{0(3)} - \widehat{\mathbf{D}}_{(3)})\mathbf{z}_i$, we obtain

$$\sum_i \|(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i\|^2 \le 2\sum_i (\boldsymbol{\epsilon}_i + \mathbf{r}_i)^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i, \tag{7}$$

Next we bound $\sum_i \boldsymbol{\epsilon}_i^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i = \sum_i \boldsymbol{\epsilon}_i^{\mathrm{T}}(\mathbf{z}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathrm{vec}(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})$. Let $\mathbf{Z}_{nq \times pKq} = (\mathbf{z}_1 \otimes \mathbf{I}_q, \ldots, \mathbf{z}_n \otimes \mathbf{I}_q)^{\mathrm{T}}$ and define the set, with any fixed $\mathbf{Z}$, $\Gamma(\mathbf{Z}) = \{\boldsymbol{\eta} = \mathbf{Z}\mathbf{d}/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})} : \mathbf{d} = \mathrm{vec}(\mathbf{D}_{(3)}), \|\mathbf{D}\| \le 1, \mathbf{D} \text{ satisfies (3)}\}$, where $\lambda_{\max}(.)$ denotes the largest eigenvalue of a symmetric matrix.

We compute the covering number of $N(\epsilon, \Gamma(\mathbf{Z}), l_2)$. Let $\mathcal{D}_1 = \{\mathbf{d} : \mathbf{d} = \mathrm{vec}(\mathbf{D}_{(3)}), \|\mathbf{D}\| \le 1, \mathbf{D} \text{ satisfies (3)}\}$. Obviously, since $\|\mathbf{Z}(\mathbf{d}_1 - \mathbf{d}_2)/\sqrt{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}\| \le \|\mathbf{d}_1 - \mathbf{d}_2\|$, $N(\epsilon, \Gamma(\mathbf{Z}), l_2) \le N(\epsilon, \mathcal{D}_1, l_2)$ and the latter is nonrandom. For $\mathbf{D} = \mathbf{S} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$ with $\mathbf{A}, \mathbf{B}, \mathbf{C}$ orthogonal, using (2) we have $\|\mathbf{D}\| = \|\mathbf{S}\|$. The ϵ-covering number for $\{\mathbf{S} : \|\mathbf{S}\| \le 1\}$ is bounded by $(C/\epsilon)^{r_1 r_2 r_3}$. For orthogonal matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, we use the distance $\|\mathbf{A}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}\|_{op}$, where $\|.\|_{op}$ denotes the operator norm. By Proposition 8 of Szarek (1982) the covering number of the set of $p \times r_1$ orthogonal matrices under this distance is bounded by $(C/\epsilon)^{r_1(p-r_1)}$. Similarly the covering number for orthogonal matrices of dimension as that of $\mathbf{B}$ and $\mathbf{C}$ can

be obtained. We have

$$\begin{aligned} & \|\mathbf{S}_1 \times_1 \mathbf{A}_1 \times_2 \mathbf{B}_1 \times_3 \mathbf{C}_1 - \mathbf{S}_2 \times_1 \mathbf{A}_2 \times_2 \mathbf{B}_2 \times_3 \mathbf{C}_2\| \\ = \; & \|(\mathbf{S}_1 - \mathbf{S}_2) \times_1 \mathbf{A}_1 \times_2 \mathbf{B}_1 \times_3 \mathbf{C}_1\| \\ & + \|\mathbf{S}_2 \times_1 (\mathbf{A}_1 - \mathbf{A}_2) \times_2 \mathbf{B}_1 \times_3 \mathbf{C}_1\| \\ & + \|\mathbf{S}_2 \times_1 \mathbf{A}_2 \times_2 (\mathbf{B}_1 - \mathbf{B}_2) \times_3 \mathbf{C}_1\| \\ & + \|\mathbf{S}_2 \times_1 \mathbf{A}_2 \times_2 \mathbf{B}_2 \times_3 (\mathbf{C}_1 - \mathbf{C}_2)\|. \end{aligned}$$

Using (2), it is easily seen that $\|(\mathbf{S}_1 - \mathbf{S}_2) \times_1 \mathbf{A}_1 \times_2 \mathbf{B}_1 \times_3 \mathbf{C}_1\| = \|\mathbf{S}_1 - \mathbf{S}_2\|$, $\|\mathbf{S}_2 \times_1 (\mathbf{A}_1 - \mathbf{A}_2) \times_2 \mathbf{B}_1 \times_3 \mathbf{C}_1\| = \|(\mathbf{A}_1 - \mathbf{A}_2)\mathbf{S}_{(1)}(\mathbf{C}_1 \otimes \mathbf{B}_1)^{\mathrm{T}}\| \leq \|\mathbf{A}_1 - \mathbf{A}_2\|_{op}$ since $\|\mathbf{S}_{(1)}\| \leq 1$ and $\|\mathbf{C}_1 \otimes \mathbf{B}_1\|_{op} = 1$, and thus

$$N(\epsilon, \Gamma(\mathbf{Z}), l_2) \leq N(\epsilon, \mathcal{D}_1, l_2) \leq (C/\epsilon)^{df},$$

where $df = r_1 r_2 r_3 + p r_1 + K r_2 + q r_3 - r_1^2 - r_2^2 - r_3^2$.

Furthermore, since the error is sub-Gaussian, we have

$$E[\exp\{a\bar{\boldsymbol{\epsilon}}^{\mathrm{T}}\boldsymbol{\eta}\}|\mathbf{Z}] \leq \exp\{Ca^2\|\boldsymbol{\eta}\|^2\},$$

where $\bar{\boldsymbol{\epsilon}} = (\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_n^{\mathrm{T}})^{\mathrm{T}}$, for any $\boldsymbol{\eta} \in \mathbb{R}^{nq}$ that can depend on predictors, in particular for $\boldsymbol{\eta} \in \Gamma(\mathbf{Z})$.

Using Corollary 2.2.8 of van der Vaart and Wellner (1996), we get

$$E[\sup_{\boldsymbol{\eta} \in \Gamma} \boldsymbol{\eta}^{\mathrm{T}}\bar{\boldsymbol{\epsilon}}] \leq C \int_0^2 \sqrt{df \log(\frac{C}{\epsilon})} d\epsilon \leq C\sqrt{df}.$$

The above implies that

$$\sum_i \boldsymbol{\epsilon}_i^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i = \sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}\|\widehat{\mathbf{D}} - \mathbf{D}_0\| O_p(\sqrt{df}).$$

Furthermore, by the Cauchy-Schwarz inequality,

$$\begin{aligned} & \sum_i \mathbf{r}_i^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i \\ \leq \; & (\sum_i \|\mathbf{r}_i\|^2)^{1/2} (\sum_i \|(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i\|^2)^{1/2}. \end{aligned}$$

Using the above two displays in (7), we get

$$\sum_i \|(\widehat{\mathbf{D}}_{(3)} - \mathbf{D}_{0(3)})\mathbf{z}_i\|^2 = O_p(df + \sum_i \|\mathbf{r}_i\|^2),$$

and thus

$$\|\widehat{\mathbf{D}} - \mathbf{D}_0\|^2 = O_p(df/n + \sum_i \|\mathbf{r}_i\|^2/n)$$

The proof for the case without intercept is complete since

$$\sqrt{\sum_i \|\mathbf{r}_i\|^2} = O_p(\sqrt{n} pq K^{-d}).$$

Now we describe the modifications required for incorporating intercepts. In this case we can incorporate $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^{\mathrm{T}}$ into $\mathbf{D}_{(3)}\mathbf{z}_i$ so that $\boldsymbol{\mu}$ is the first column of $\mathbf{D}_{(3)}$ while $\mathbf{z}_i$ is replaced by $(1, \mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}}$. The covering number with this expanded set of parameters has the same order since only q parameters are added and $q = O(df)$. Thus the previous arguments can still be followed exactly to obtain the same convergence rate.

Finally, to get the bound in the approximate low-rank case, we assume the existence of $\widetilde{\mathbf{D}} \in \mathcal{D}(r_1, r_2, r_3)$ with $\|\widetilde{\mathbf{D}} - \mathbf{D}_0\| \leq \xi$. Starting from the basic inequality

$$\sum_i \|\mathbf{y}_i - \widehat{\mathbf{D}}_{(3)}\mathbf{z}_i\|^2 \leq \sum_i \|\mathbf{y}_i - \widetilde{\mathbf{D}}_{(3)}\mathbf{z}_i\|^2,$$

we get, similar to (7),

$$\sum_i \|(\widehat{\mathbf{D}}_{(3)} - \widetilde{\mathbf{D}}_{(3)})\mathbf{z}_i\|^2 \leq 2\sum_i (\boldsymbol{\epsilon}_i + \mathbf{r}_i)^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \widetilde{\mathbf{D}}_{(3)})\mathbf{z}_i,$$

where $\mathbf{r}_i = (r_{i1}, \ldots, r_{iq})^{\mathrm{T}}$ with $r_{il} = \sum_j f_{jl}(x_{ij}) - \sum_{j,k} \widetilde{d}_{jkl} b_k(x_{ij})$. Decompose $\mathbf{r}_i = \mathbf{r}_{i1} + \mathbf{r}_{i2}$, $\mathbf{r}_{i1} = (r_{i11}, \ldots, r_{i1q})^{\mathrm{T}}$, $\mathbf{r}_{i2} = (r_{i21}, \ldots, r_{i2q})^{\mathrm{T}}$, $r_{i1l} = \sum_j f_{jl}(x_{ij}) - \sum_{j,k} d_{0jkl} b_k(x_{ij})$, $r_{i2l} = \sum_{j,k}(d_{0jkl} - \widetilde{d}_{jkl}) b_k(x_{ij})$. The terms $2\sum_i (\boldsymbol{\epsilon}_i + \mathbf{r}_{i1})^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \widetilde{\mathbf{D}}_{(3)})\mathbf{z}_i$ can be bounded in the same way as for the exactly low-rank case. We also have $\sum_i \mathbf{r}_{i2}^{\mathrm{T}}(\widehat{\mathbf{D}}_{(3)} - \widetilde{\mathbf{D}}_{(3)})\mathbf{z}_i = (\sum_i \|\mathbf{r}_{i2}\|^2)^{1/2}(\sum_i \|(\widehat{\mathbf{D}}_{(3)} - \widetilde{\mathbf{D}}_{(3)})\mathbf{z}_i\|^2)^{1/2}$, and $\sum_i \|\mathbf{r}_{i2}\|^2 = O_p(n\xi^2)$ using the approximate low-rank assumption and (A2). These together shows $\|\widehat{\mathbf{D}} - \widetilde{\mathbf{D}}\| = O_p(\sqrt{df/n} + pqK^{-d} + \xi)$ and in turn $\|\widehat{\mathbf{D}} - \mathbf{D}_0\| = O_p(\sqrt{df/n} + pqK^{-d} + \xi)$. □

Proof of Theorem 2. As before we ignore the intercept for simplicity of notation (or regard the intercepts as already being incorporated into $\mathbf{Zd}$ but suppressed in notation). Let $\widehat{\mathbf{D}}^o$ be the oracle estimator defined as the minimizer of (5) using only s relevant covariates and then padded with zeros to make it a $p \times K \times q$ tensor. We show that the oracle estimator herein defined is a local minimizer of (6), with probability approaching one.

Consider $\mathbf{D}$ in a neighborhood of $\widehat{\mathbf{D}}^o$ with $\|\mathbf{d}_j - \widehat{\mathbf{d}}_j^o\| \leq \delta$ for $j \leq s$ and $\|\mathbf{d}_j\| \leq \delta$ for $j > s$, for some $\delta > 0$ sufficiently small. More concretely, define a neighborhood of $\widehat{\mathbf{D}}^o$ as $\mathcal{D}_\delta = \{\mathbf{D} : \mathbf{D} \text{ satisfies (3)}, \|\mathbf{d}_j - \widehat{\mathbf{d}}_j^o\| \leq \delta \text{ for } j \leq s \text{ and } \|\mathbf{d}_j\| \leq \delta \text{ for } j > s\}$.

We only need to show that for δ small enough,

$$\|\bar{\mathbf{y}} - \mathbf{Z}\widehat{\mathbf{d}}^o\|^2 + n\sum_{j=1}^p p_\lambda(\|\mathbf{Z}_j\widehat{\mathbf{d}}_j^o\|/\sqrt{n}) \leq \inf_{\mathbf{D}\in\mathcal{D}_\delta} \|\bar{\mathbf{y}} - \mathbf{Zd}\|^2 + n\sum_{j=1}^p p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}), \tag{8}$$

with probability approaching one as $n \to \infty$.

We first note that under the assumption $\lambda^2 \ll \min_{j\leq s}\sum_{l=1}^q \|f_{jl}\|^2$, we have that $\inf_{j\leq s}\|\mathbf{Z}_j\mathbf{d}_{0j}/\sqrt{n}\| \asymp \inf_{j\leq s}\|\mathbf{d}_{0j}\| \gg \lambda$ in probability. With the convergence rate $\|\mathbf{d}_j^o - \mathbf{d}_0\| = O_p(\sqrt{df/n} + sqK^{-d}) = o_p(\lambda)$, we also have $\|\mathbf{Z}_j\widehat{\mathbf{d}}_j^o\| \gg \lambda$ with probability approaching one. In particular, we have $\|\mathbf{Z}_j\widehat{\mathbf{d}}_j^o\| > a\lambda$ for $j \leq s$ and thus if δ is small enough, $\|\mathbf{Z}_j\mathbf{d}_j\| > a\lambda$ for any $\mathbf{d} \in \mathcal{D}_\delta$, which implies $p_\lambda(\|\widehat{\mathbf{d}}_j^o\|) = p_\lambda(\|\mathbf{d}_j\|)$ for $j \leq s$. Thus, using that $\widehat{\mathbf{D}}^o$ is the oracle estimator that minimizes (5), we have

$$\|\bar{\mathbf{y}} - \mathbf{Z}\widehat{\mathbf{d}}^o\|^2 + n\sum_{j=1}^p p_\lambda(\|\mathbf{Z}_j\widehat{\mathbf{d}}_j^o\|/\sqrt{n}) \leq \|\bar{\mathbf{y}} - \mathbf{Zd}\|^2 + n\sum_{j=1}^p p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}), \tag{9}$$

for any $\mathbf{D} \in \mathcal{D}'_\delta$ where $\mathcal{D}'_\delta = \{\mathbf{D} : \mathbf{d}_j = \mathbf{0} \text{ for } j > s\} \cap \mathcal{D}_\delta$.

Thus (8) will be proved if we can show that

$$\inf_{\mathbf{D}\in\mathcal{D}'_\delta} \|\bar{\mathbf{y}} - \mathbf{Z}\mathbf{d}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}) \le \inf_{\mathbf{D}\in\mathcal{D}_\delta\setminus\mathcal{D}'_\delta} \|\bar{\mathbf{y}} - \mathbf{Z}\mathbf{d}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}).$$

For this, we will show that if δ is small enough, for any $\mathbf{d} \in \mathcal{D}_\delta$, we can construct a $\widetilde{\mathbf{d}} \in \mathcal{D}'_\delta$ such that

$$\|\bar{\mathbf{y}} - \mathbf{Z}\widetilde{\mathbf{d}}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\mathbf{Z}_j\widetilde{\mathbf{d}}_j\|/\sqrt{n}) \le \|\bar{\mathbf{y}} - \mathbf{Z}\mathbf{d}\|^2 + n\sum_{j=1}^{p} p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}). \tag{10}$$

In fact, given $\mathbf{D} \in \mathcal{D}_\delta$, we let $\widetilde{\mathbf{D}}$ be such that $\widetilde{\mathbf{d}}_j = \mathbf{d}_j$ for $j \le s$ and $\widetilde{\mathbf{d}}_j = \mathbf{0}$ for $j > s$. Then the right hand side of (10) subtracting its left hand side is equal to

$$\|\sum_{j=s+1}^{p} \mathbf{Z}_j\mathbf{d}_j\|^2 - 2\langle \bar{\mathbf{y}} - \sum_{j=1}^{s} \mathbf{Z}_j\mathbf{d}_j, \sum_{j=s+1}^{p} \mathbf{Z}_j\mathbf{d}_j\rangle + n\sum_{j=s+1}^{p} p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}). \tag{11}$$

Furthermore,

$$\begin{aligned}
&|\langle \bar{\mathbf{y}} - \sum_{j=1}^{s} \mathbf{Z}_j\mathbf{d}_j, \sum_{j=s+1}^{p} \mathbf{Z}_j\mathbf{d}_j\rangle| \\
\le\ & |\sum_{j=s+1}^{p} \langle \mathbf{Z}_j^{\mathrm{T}}(\bar{\mathbf{y}} - \sum_{j'=1}^{s} \mathbf{Z}_{j'}\widehat{\mathbf{d}}^o_{j'}), \mathbf{d}_j\rangle| + |\sum_{j=s+1}^{p} \langle \mathbf{Z}_j^{\mathrm{T}}(\sum_{j'=1}^{s} \mathbf{Z}_{j'}(\widehat{\mathbf{d}}^o_{j'} - \mathbf{d}_{j'})), \mathbf{d}_j\rangle| \\
\le\ & \max_{j>s} \|\mathbf{Z}_j^{\mathrm{T}}(\bar{\mathbf{y}} - \sum_{j'=1}^{s} \mathbf{Z}_{j'}\widehat{\mathbf{d}}^o_{j'})\| \left(\sum_{j>s} \|\mathbf{d}_j\|\right) + \sum_{j>s, j'\le s} \|\mathbf{Z}_j\|\|\mathbf{Z}_{j'}\|\|\widehat{\mathbf{d}}^o_{j'} - \mathbf{d}_{j'}\|\|\mathbf{d}_j\| \\
\le\ & \left(\max_{j>s} \|\mathbf{Z}_j^{\mathrm{T}}\bar{\boldsymbol{\epsilon}}\| + \max_{j>s} \|\mathbf{Z}_j^{\mathrm{T}}\bar{\mathbf{r}}\| + \max_{j>s} \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}(\widehat{\mathbf{d}}^o - \mathbf{d}_0)\|\right)(\sum_{j>s} \delta_j) + (\sum_{j>s, j'\le s} \|\mathbf{Z}_j\|\|\mathbf{Z}_{j'}\|\delta_j\delta_{j'}) \\
\le\ & \left(\max_{j>s} \|\mathbf{Z}_j^{\mathrm{T}}\bar{\boldsymbol{\epsilon}}\| + O_p(\sqrt{df} + \sqrt{n}sqK^{-d})\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}\right)(\sum_{j>s} \delta_j) + (\sum_{j>s, j'\le s} \|\mathbf{Z}_j\|\|\mathbf{Z}_{j'}\|\delta_j\delta_{j'}),
\end{aligned} \tag{12}$$

where $\bar{\mathbf{r}} = (\mathbf{r}_1^{\mathrm{T}}, \ldots, \mathbf{r}_n^{\mathrm{T}})^{\mathrm{T}}$ and we denote $\delta_j = \|\widehat{\mathbf{d}}^o_j - \mathbf{d}_j\|$. By the definition of $\mathbf{Z}_j$, we have

$$\max_j \|\mathbf{Z}_j^{\mathrm{T}}\bar{\boldsymbol{\epsilon}}\| \le \sqrt{Kq}\max_{j,k,l} |\sum_i b_k(x_{ij})\epsilon_{il}|.$$

By the sub-Gaussianity of the error, and conditional on predictors, we have by Lemma 2.2.2 of van der Vaart and Wellner (1996),

$$E[\max_{j,k,l} |\sum_i b_k(x_{ij})\epsilon_{il}| |\{\mathbf{x}_i\}] \le C\sqrt{\log(pKq)}\max_{j,k,l} \sqrt{\sum_i b_k^2(x_{ij})}.$$

We have $E[b_k^2(x_{ij})]$ is bounded. Also, $E[|b_k^2(x_{ij})|^r] \leq Cr!K^{r-1}$. Then by Bernstein's inequality,

$$P(\sum_i b_k^2(x_{ij}) - E\sum_i b_k^2(x_{ij}) > Cn) \leq 2\exp\{-Cn/K\},$$

and by the union bound we will get

$$\max_{j,k,l} \sqrt{\sum_i b_k^2(x_{ij})} = O_p(\sqrt{n}),$$

if $K\log(pKq)/n \to 0$.

Now, if δ is small enough, for the SCAD penalty, $p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}) = \lambda\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n} \asymp \lambda\|\mathbf{d}_j\|$ by the definition of the penalty. For the MCP, we also have $p_\lambda(\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n}) \geq (\lambda/2)\|\mathbf{Z}_j\mathbf{d}_j\|/\sqrt{n} \asymp \lambda\|\mathbf{d}_j\|$. Thus, (11) is larger than

$$(Cn\lambda - O_p(\sqrt{n\log(pKq)} + \sqrt{ndf} + nsqK^{-d})\sum_{j>s}\delta_j - (\sum_{j>s,j'\leq s}\|\mathbf{Z}_j\|\|\mathbf{Z}_{j'}\|\delta_j\delta_{j'})$$

Obviously, when the scalar in front of $\sum_{j>s}\delta_j$ above is positive, the first term dominates the second term if δ_j is sufficiently small, which proved the theorem. □

Appendix: Table of notations

Table 5: Table of notations.

n	sample size
K	number of basis functions for splines
p	number of predictors
q	number of responses
$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$	observed predictor values for the i-th individual, $i \in \{1, \ldots, n\}$
$\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})^{\mathrm{T}}$	observed response values for the i-th individual, $i \in \{1, \ldots, n\}$
$\mathbf{b}(x) = (b_1(x), \ldots, b_K(x))^{\mathrm{T}}$	spline basis functions
$\mathbf{T}_{(n)}$	matricization of a tensor $\mathbf{T}$ along mode n
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^{\mathrm{T}}$	intercept in the additive model
f_{jl}	unknown functions to be estimated in the additive model, $j \in \{1, \ldots, p\}, l \in \{1, \ldots, q\}$
$\mathbf{d}_{jl} = (d_{j1l}, \ldots, d_{jKl})^{\mathrm{T}}$	spline coefficients for the approximation of f_{jl}
$\mathbf{d}_j = (\mathbf{d}_{j1}^{\mathrm{T}}, \ldots, \mathbf{d}_{jK}^{\mathrm{T}})^{\mathrm{T}}$	all spline coefficients associated with predictor j arranged as a vector
$\mathbf{D}$	the $p \times K \times q$ tensor containing all d_{jkl}
(r_1, r_2, r_3)	rank for the low-rank approximation of $\mathbf{D}$
$\mathcal{D}(r_1, r_2, r_3)$	the set of all tensors of dimension $p \times K \times q$ with rank (r_1, r_2, r_3)
$\mathbf{z}_i$	the pK-dimensional vector with components $b_k(x_{ij})$
$\mathbf{Z}$, $\mathbf{Z}_j$	The $(nq) \times (pKq)$ matrix $(\mathbf{z}_1 \otimes \mathbf{I}_q, \ldots, \mathbf{z}_n \otimes \mathbf{I}_q)^{\mathrm{T}}$, with $\mathbf{Z}_j$ the submatrix containing columns associated with predictor j

References

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1. JMLR Workshop and Conference Proceedings, 2012.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014a.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv preprint arXiv:1402.5180*, 2014b.

T W Anderson. Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27(4):1141–1154, 1999. ISSN 0090-5364.

P Breheny and J Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25:173–187, 2015.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011. ISSN 0090-5364.

Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007. ISSN 0090-5364.

Kun Chen, Kung Sik Chan, and Nils Chr Stenseth. Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):203–221, 2012. ISSN 1467-9868.

Kun Chen, Hongbo Dong, and Kung Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

Koei Chin, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, et al. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, 10(6): 529–541, 2006.

L de Lathauwer, B de Moor, and J Vandewalle. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000. ISSN 0895-4798. doi: 10.1137/S0895479896305696.

J Q Fan and R Z Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. ISSN 0162-1459.

Jianqing Fan, Yunbei Ma, and Wei Dai. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284, 2014. ISSN 0162-1459.

LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. ISSN 0040-1706.

John Geweke. Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75(1):121–146, 1996. ISSN 0304-4076.

Botao Hao, Boxiang Wang, Pengyuan Wang, Jingfei Zhang, Jian Yang, and Will Wei Sun. Sparse tensor additive regression. *Journal of machine learning research*, 22:1–43, 2021.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York, 1st edition, 1990. ISBN 0412343908.

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

Shiyuan He, Kejun He, and Jianhua Z. Huang. Improved estimation of high-dimensional additive models using subspace learning. *Journal of Computational and Graphical Statistics*, 31(3):866–876, 2022.

X He and P Shi. Bivariate tensor-product B-splines in a partly linear model. *Journal of Multivariate Analysis*, 58(2):162–181, 1996. ISSN 0047-259X.

J L Horowitz and S Lee. Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249, 2005. ISSN 0162-1459.

J Huang, J L Horowitz, and F Wei. Variable selection in nonparametric additive models. *Annals of Statistics*, 38(4):2282–2313, 2010.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. ISSN 0047-259X.

Majid Janzamin, Rong Ge, Jean Kossaifi, Anima Anandkumar, et al. Spectral learning on matrices and tensors. *Foundations and Trends® in Machine Learning*, 12(5-6):393–536, 2019.

Qian Jiang, Hansheng Wang, Yingcun Xia, and Guohua Jiang. On a principal varying coefficient model. *Journal of the American Statistical Association*, 108(501):228–236, 2013. ISSN 0162-1459.

Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, 2009. ISSN 0036-1445. doi: 10.1137/07070111X.

Heng Lian and Hua Liang. Generalized additive partial linear models with high-dimensional covariates. *Econometric Theory*, 29:1136–1161, 2013. ISSN 1469-4360.

Oliver Linton and Jens Perch Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100, 1995. ISSN 0006-3444.

X Liu, L Wang, and H Liang. Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21:1225–1248, 2011.

L Meier, S de Geer, and P Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37(6B):3779–3821, 2009.

Michelle Miranda, Hongtu Zhu, and Joseph G Ibrahim. TPRM: Tensor partition regression models with applications in imaging biomarker detection. *arXiv preprint arXiv:1505.05482*, 2015.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011. ISSN 0090-5364.

Garvesh Raskutti, Ming Yuan, and H Chen. Convex regularization for high-dimensional multi-response tensor regression. *arXiv:1512.01215*, 2017.

P Ravikumar, H Liu, J Lafferty, and L Wasserman. SpAM: Sparse additive models. *Advances in Neural Information Processing Systems 20*, pages 1201–1208, 2008.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Larry Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007. ISBN 1139463438.

Hanie Sedghi and Ashish Sabharwal. Knowledge completion for generics using guided tensor factorization. *Transactions of the Association for Computational Linguistics*, 6:197–210, 2018.

Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics*, pages 1223–1231. PMLR, 2016.

Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13:689–705, 1985. ISSN 0090-5364.

Charles J. Stone. The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14(2):590–606, 1986. ISSN 0090-5364. doi: 10.1214/aos/1176349940.

W Sun and L Li. STORE: sparse tensor response regression and neuroimaging analysis. *Journal of Machine Learning Research*, 18:1–37, 2017.

Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society Series B-Methodological*, to appear, 2017.

Stanisław J Szarek. Nets of Grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa*, pages 169–185, 1982.

R Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288, 1996. ISSN 0035-9246.

A B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009. ISBN 0387790519.

A W van der Vaart and J A Wellner. *Weak convergence and empirical processes*. Springer Verlag, 1996. ISBN 0387946403.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

L Wang, X Liu, H Liang, and R J Carroll. Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics*, 39(4):1827–1851, 2011. ISSN 0090-5364.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

L Xue. Consistent variable selection in additive models. *Statistica Sinica*, 19:1281–1296, 2009.

C H Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. ISSN 0090-5364.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society Series B-Methodological*, 76:463–483, 2014.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor Regression with Applications in Neuroimaging Data Analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013. ISSN 0162-1459.

H Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. ISSN 0162-1459.