Statistica Sinica Preprint No: SS-2022-0103	
Title	Multivariate Varying-coefficient Models via Tensor Decomposition
Manuscript ID	SS-2022-0103
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0103
Complete List of Authors	Fengyu Zhang, Ya Zhou, Kejun He and Raymond K. W. Wong
Corresponding Authors	Kejun He
E-mails	kejunhe@ruc.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Multivariate Varying-coefficient Models via Tensor Decomposition

Fengyu Zhang, Ya Zhou, Kejun He, and Raymond K. W. Wong

Renmin University of China, Chinese Academy of Medical Sciences and Peking Union Medical College, Renmin University of China, and Texas A&M University

Abstract: Multivariate varying-coefficient models (MVCM) are popular statistical tools for analyzing the relationship between multiple responses and covariates. Nevertheless, estimating large numbers of coefficient functions is challenging, especially with a limited amount of samples. In this work, we propose a reduced-dimension model based on the Tucker decomposition, which unifies several existing models. In addition, sparse predictor effects, in the sense that only a few predictors are related to the responses, are exploited to achieve an interpretable model and sufficiently reduce the number of unknown functions to be estimated. All the above dimension-reduction and sparsity considerations are integrated into a penalized least squares problem on the constraint domain of 3rd-order tensors. To compute the proposed estimator, we propose a block updating algorithm with ADMM and manifold optimization. We also establish the oracle inequality for the prediction risk of the proposed estimator. A real data set from Framingham Heart Study is used to demonstrate the good predictive performance of the proposed method.

Key words and phrases: Tucker low rank; group Lasso; dimensionality reduction; polynomial splines; sparsity.

The first two authors contributed equally to this work. Kejun He (Email: kejunhe@ruc.edu.cn) is the corresponding author.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

1. Introduction

Varying coefficient models (VCMs, Hastie and Tibshirani, 1993) are among the popular class of structured regression models, which have reasonably flexible nonparametric components yet can be estimated well with a moderate amount of data (Ruppert et al., 2003). In VCMs, the regression coefficients of predictors vary with an observable exposure variable. VCMs have been extensively studied in literature and widely used in practice; see, for examples, Hoover et al. (1998), Huang et al. (2002), Park et al. (2015), and the references therein. For settings with a large number of predictors (possibly larger than the sample size), Wang et al. (2008) adopted basis function expansions and the SCAD penalty to address the problem of variable selection. Wei et al. (2011) and Lian (2012) applied adaptive group Lasso and spline function approximations to simultaneously identify relevant predictors and estimate varying coefficient functions of the selected ones. For their estimators, these papers obtained the rate of convergence and variable selection consistency under suitable conditions. Besides, Xue and Qu (2012) utilized truncated ℓ_1-penalty (TLP) to select variables and obtained the oracle properties for their varying-coefficient estimator. To enhance the computational scalability, feature screening techniques for the VCM were considered in Fan et al. (2014) and Liu et al. (2014) by ranking some proposed measures of the marginal nonparametric contributions of each predictor given the exposure variable, and the sure independent screening properties were investigated.

In many applications, multiple responses are jointly observed with the predictors and exposure variable. For instance, the Framingham Heart Study (Dawber et al., 1951) collected multiple phenotype variables from each patient to identify the common factors related to cardiovascular diseases. Obviously, one can simply model each response variable separately using VCMs. These

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

models are together viewed as a regression model for the multivariate response, called (unstructured) multivariate varying coefficient model (MVCM). One challenge associated with such a MVCM is the significant number of coefficient functions required to be estimated. More specifically, we need to estimate pq functions, if there are p covariates and q response variables. To circumvent this problem, structures among these pq functions should be exploited. He et al. (2018) proposed a principal-component-based approach by assuming all the coefficient functions can be approximated by linear combinations of a much smaller number of unknown functions. But the authors did not exploit the correlation between the responses and their method cannot handle the settings with large number of response variables. Lian and Ma (2013), on the other hand, assumed a low-rank structure in the conditional means of the responses among the samples. However, their model does not take into account the correlations among the predictors and/or the varying coefficients. Further, they did not propose an efficient algorithm to solve their penalized least squares problem.

In this work we propose a novel method based on dimension-reduction tools for tensors (Kolda and Bader, 2009) to handle the MVCM under high-dimensional settings. In particular, we show that dimension reductions in the predictors, the space of coefficient functions, and the responses correspond to the low rankness in the first, second, and third mode of a 3rd-order tensor. We thus propose to use the idea of Tucker decomposition (Tucker, 1966) to integrate these three dimension reductions into a simple notion of low multilinear rank. Both the work of He et al. (2018) and Lian and Ma (2013) can be treated as special cases of our proposed model. In addition, sparse predictor effects, in the sense that only a few of predictors are related to the responses, is often a reasonable assumption in high-dimensional settings. All the above dimension-reduction and

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

sparsity considerations can be incorporated into the estimation procedure through a penalized least squares problem on the constraint domain of 3rd-order tensors. To compute the proposed estimator, we design a block updating algorithm with ADMM (Boyd et al., 2011) and manifold optimization (Edelman et al., 1998; Absil et al., 2009). We also establish the oracle inequality for the prediction risk of the proposed estimator.

The rest of the paper is organized as follows. In Section 2 we introduce the proposed reduced multivariate varying-coefficient model using Tucker decomposition. The estimation method and computational details are presented in Sections 3 and 4, respectively. We establish the oracle inequality for the prediction risk of the proposed estimator in Section 5. We use both a simulation study and a real data application in Section 6 to illustrate the practical performance of the proposed method. The main contributions of this paper are summarized in Section 7 with some concluding remarks. Technical details are provided in a separate online supplemental document.

2. Model

Let $\boldsymbol{y} = (y_1, \ldots, y_q)^\intercal$, $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$, and t be the q-dimensional vector of responses, the p-dimensional vector of predictors, and the exposure variable with compact domain $\mathcal{T}$, respectively. Without loss of generality, we assume $\mathcal{T} = [0, 1]$. Each response is posited to follow the univariate-response VCM, i.e.,

$$y_l = \sum_{j=1}^{p} f_{jl}(t) x_j + \epsilon_l, \quad l = 1, \ldots, q, \tag{2.1}$$

where $f_{jl}(t)$'s are the coefficient functions and, ϵ_l's are the noise variables with mean 0 and variance σ_l^2. These noise variables are independent of $(\boldsymbol{x}, t)$. By setting $x_1 = 1$, we can see that the model can accommodate an intercept function. In vector-matrix notation, (2.1) can be

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

written as

$$\boldsymbol{y} = \boldsymbol{F}(t)^\intercal \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{F}(t) = (f_{jl}(t))_{p\times q}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_q)^\intercal$. We call (2.2) the full model of MVCM, where there are in total pq varying coefficient functions to be estimated nonparametrically.

When pq is relatively large, there are huge numbers of nonparametric functions, which are difficult to estimate accurately with a small or moderate amount of data. To cope with this challenge, Lian and Ma (2013) assumed a rank-R_3 structure on the matrix of coefficient functions with $R_3 < q$, aiming to reduce the model complexity among the responses. Specifically, Lian and Ma (2013) proposed to reduce the full MVCM (2.2) to

$$\boldsymbol{y} = \boldsymbol{C}\widetilde{\boldsymbol{F}}(t)^\intercal \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\boldsymbol{C} \in \mathbb{R}^{q\times R_3}$ with $\boldsymbol{C}^\intercal\boldsymbol{C} = \boldsymbol{I}_{R_3}$ and $\widetilde{\boldsymbol{F}}(t)$ is a matrix of $p \times R_3$ unknown functions. Model (2.3) implies that the means of the responses conditional on the predictors and exposure variable are assumed to be R_3 linearly dependent among the samples. Compared with (2.2), the number of parameters is reduced to pR_3 functions together with a $q \times R_3$ coefficient matrix. He et al. (2018), on the other hand, proposed a functional principal-component-based approach which assumes all pq coefficient functions can be well approximated by a small number of R_2 unknown data-driven principal functions $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_{R_2}(t))^\intercal$. More precisely, they assumed the vectorized $\boldsymbol{F}(t)$ can be represented by $\mathrm{vec}\{\boldsymbol{F}(t)\} = \boldsymbol{D}\boldsymbol{\beta}(t)$ with a coefficient matrix $\boldsymbol{D} \in \mathbb{R}^{pq\times R_2}$, and the conditional mean of responses in the full MVCM (2.2) is then reduced to

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}, t) = \mathrm{vec}\{\boldsymbol{x}^\intercal\boldsymbol{F}(t)\} = (\boldsymbol{I}_q \otimes \boldsymbol{x}^\intercal)\mathrm{vec}\{\boldsymbol{F}(t)\} = (\boldsymbol{I}_q \otimes \boldsymbol{x}^\intercal)\boldsymbol{D}\boldsymbol{\beta}(t). \tag{2.4}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

For model identifiability, the principal functions $\boldsymbol{\beta}(t)$ are required to be orthonormal, i.e.,

$$\int_{\mathcal{T}} \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^{\intercal}\,\mathrm{d}t = \boldsymbol{I}_{R_2}.$$

Thus, one only needs to estimate R_2 principal functions together with a $p \times R_2 \times q$ coefficient tensor for a reduced MVCM in (2.4). In the univariate-response VCM, i.e., $q = 1$, Jiang et al. (2013) proposed another principal component varying coefficient model. Specifically, treating the l-th response in (2.1) as a single response, the model of Jiang et al. (2013) is equivalent to

$$y_l = \boldsymbol{f}_l(t)^{\intercal}\boldsymbol{A}^{\intercal}\boldsymbol{x}^{\intercal} + \epsilon_l, \tag{2.5}$$

where $\boldsymbol{f}_l(t)$ is a vector of R_1 unknown functions and $\boldsymbol{A} \in \mathbb{R}^{p\times R_1}$ is the principal loading matrix. Overall, Models (2.3), (2.4), and (2.5) encompass dimension reductions within the responses, the coefficient functions, and the predictors, respectively.

However, the above models are hard to compare since each model focuses on a different way of dimension reduction. In this work, we observe that these models can be unified into a general model, which allows simultaneous reductions and provides a coherent understanding of these methods. To illustrate this idea, we begin with the form of (2.4). Denote $\bar{\boldsymbol{S}} \in \mathbb{R}^{p\times R_2\times q}$ to be a 3rd-order tensor satisfying $\bar{\boldsymbol{S}}_{(2)} = \boldsymbol{D}^{\intercal}$. Model (2.4) can be written as

$$\boldsymbol{y} = \{\bar{\boldsymbol{S}} \,\bar{\times}_2\, \boldsymbol{\beta}(t)\}^{\intercal}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{2.6}$$

where $\bar{\times}_2$ denotes the 2-mode (vector) product of a tensor with a vector (Kolda and Bader, 2009). More precisely, the result of the d-mode (vector) product of a generic Nth-order tensor $\mathfrak{G} = (g_{i_1,i_2,\ldots,i_N}) \in \mathbb{R}^{I_1\times I_2\times\cdots I_N}$ and a vector $\boldsymbol{v} \in \mathbb{R}^{I_d}$ is a tenor of order $N-1$ with dimension $I_1 \times \cdots \times I_{d-1} \times I_{d+1} \times \cdots \times I_N$ such that its $(i_1, \ldots, i_{d-1}, i_{d+1}, \ldots, i_N)$-th element is $\sum_{i_d=1}^{I_d} v_{i_d} \cdot g_{i_1,i_2,\ldots,i_N}$. Such

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

a reformulation shows that exploring the correlations among the varying coefficients is equivalent to the dimension reduction on the second mode of a 3rd-order tensor. Figure 1 illustrates the corresponding matrix of coefficient functions in (2.6) using this tensor-vector product. Similarly, the correlations among the predictors and responses are related to dimension reductions on the first and third modes, respectively.

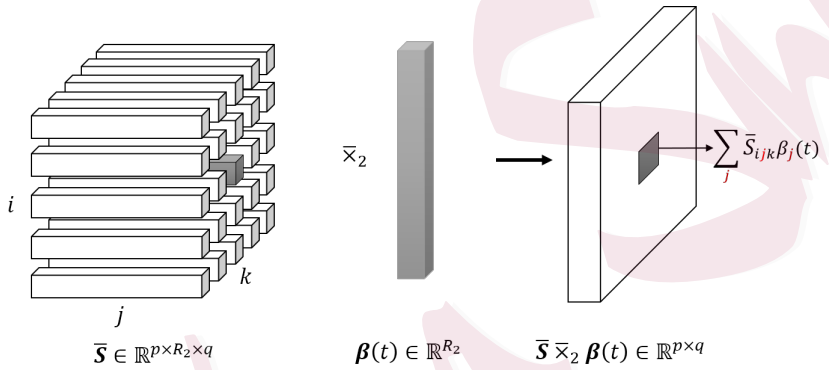

Figure 1: An illustration plot of the coefficient functions matrix in (2.6) using a tensor formulation and the 2-mode (vector) product.

Therefore, to simultaneously explore all reductions, we propose

$$\boldsymbol{y} = \{\boldsymbol{S} \times_1 \boldsymbol{A} \times_3 \boldsymbol{C} \,\bar{\times}_2\, \boldsymbol{\beta}(t)\} \,\bar{\times}_1\, \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{2.7}$$

where $\times_d$ denotes the d-mode (matrix) product of a tensor with a matrix (Kolda and Bader, 2009), $d = 1, 2, 3$; $\boldsymbol{\beta}(t)$ is a vector of R_2 unknown principal functions; $\boldsymbol{S} \in \mathbb{R}^{R_1\times R_2\times R_3}$, $\boldsymbol{A} \in \mathbb{R}^{p\times R_1}$, and $\boldsymbol{C} \in \mathbb{R}^{q\times R_3}$ are coefficients to be estimated. We depict $\boldsymbol{S} \times_1 \boldsymbol{A}$ in Figure 2 to illustrate the d-mode (matrix) product of a tensor with a matrix. Similar to Jiang et al. (2013), Lian and Ma (2013), and He et al. (2018), we require $\boldsymbol{A}$, $\boldsymbol{C}$, and $\boldsymbol{\beta}(t)$ are orthonormal, i.e.,

$$\boldsymbol{A}^\intercal\boldsymbol{A} = \boldsymbol{I}_{R_1}, \quad \boldsymbol{C}^\intercal\boldsymbol{C} = \boldsymbol{I}_{R_3}, \quad \text{and} \quad \int_{\mathcal{T}} \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^\intercal \,\mathrm{d}t = \boldsymbol{I}_{R_2}. \tag{2.8}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

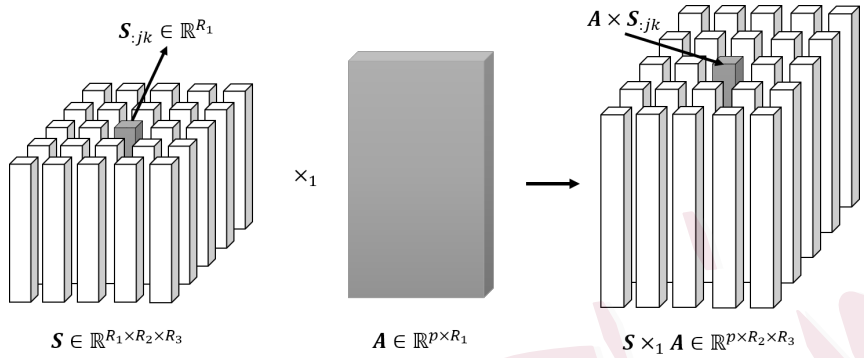

Figure 2: An illustration plot of the d-mode (matrix) product of a tensor and a matrix.

The multilinear structure of the varying coefficients $\boldsymbol{S}\times_1\boldsymbol{A}\times_3\boldsymbol{C}\ \bar{\times}_2\boldsymbol{\beta}(t)$ coincides with the Tucker decomposition (Tucker, 1966) for a 3rd-order tensor. We observe that Models (2.3), (2.4), and (2.5) are all special cases of Model (2.7). In particular, removing the first and second mode reductions in (2.7) and writing $\boldsymbol{S}\times_1\boldsymbol{A}\ \bar{\times}_2\boldsymbol{\beta}(t)=\widetilde{\boldsymbol{F}}(t)$, (2.7) can recover (2.3). Further, (2.4) can be directly obtained by letting $\bar{\boldsymbol{S}}=\boldsymbol{S}\times_1\boldsymbol{A}\times_3\boldsymbol{C}$. Finally, singling out $\boldsymbol{A}$ and treating $q=1$ in $\boldsymbol{S}\times_3\boldsymbol{C}\ \bar{\times}_2\boldsymbol{\beta}(t)$ recovers (2.5). Therefore, each mode in the decomposition $\boldsymbol{S}\times_1\boldsymbol{A}\times_3\boldsymbol{C}\ \bar{\times}_2\boldsymbol{\beta}(t)$ corresponds to one reduced model mentioned above.

We conclude this section by a remark that the constraint (2.8) does not guarantee the identifiability of the proposed model (2.7). Indeed for any $\boldsymbol{U}\in\mathbb{R}^{R_2\times R_2}$ with $\boldsymbol{U}\boldsymbol{U}^\intercal=\boldsymbol{I}_{R_2}$, we have

$$\{\boldsymbol{S}\times_1\boldsymbol{A}\times_3\boldsymbol{C}\ \bar{\times}_2\boldsymbol{\beta}(t)\}^\intercal\boldsymbol{x}=\left[(\boldsymbol{S}\times_2\boldsymbol{U})\times_1\boldsymbol{A}\times_3\boldsymbol{C}\ \bar{\times}_2\{\boldsymbol{U}\boldsymbol{\beta}(t)\}\right]^\intercal\boldsymbol{x}.$$

In other words, $(\boldsymbol{S},\boldsymbol{A},\boldsymbol{C},\boldsymbol{\beta}(t))$ and $(\boldsymbol{S}\times_2\boldsymbol{U},\boldsymbol{A},\boldsymbol{C},\boldsymbol{U}\boldsymbol{\beta}(t))$ result in the same reduced MVCM model. However, only the identification of the regression coefficient functions $\boldsymbol{F}(t)$ is needed to understand the reduced MVCM (2.7), which is fulfilled since $\boldsymbol{F}(t)=\boldsymbol{S}\times_1\boldsymbol{A}\times_3\boldsymbol{C}\ \bar{\times}_2\boldsymbol{\beta}(t)$. As for computation, these identifiability issues may lead to algorithmic instability and so some further regularizations are introduced on $(\boldsymbol{S},\boldsymbol{A},\boldsymbol{C},\boldsymbol{\beta}(t))$ in Section 3 to obtain an efficient algorithm.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

3. Penalized Least Squares Estimation

To estimate the parameters in our reduced MVCM (2.7), we first approximate the principal component functions $\boldsymbol{\beta}(t)$ using splines. Specifically, let $\boldsymbol{b}(t) = (b_1(t), \ldots, b_K(t))^\intercal$ be a vector of orthonormal B-spline basis functions with dimension K. For the r_2-th principal component function $\beta_{r_2}(t)$, we write

$$\beta_{r_2}(t) \approx \sum_{k=1}^{K} B_{k,r_2} b_k(t),$$

where B_{k,r_2}'s are the corresponding spline coefficients. Denote $\boldsymbol{B}_{r_2} = (B_{1,r_1}, \ldots, B_{K,r_2})^\intercal$. We stack $\boldsymbol{B}_{r_2}$, $r_2 = 1, \ldots, R_2$, into a matrix of coefficients, and let $\boldsymbol{B} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{R_2}) \in \mathbb{R}^{K \times R_2}$. Moreover, we require $\boldsymbol{B}$ to satisfy the constraint $\boldsymbol{B}^\intercal \boldsymbol{B} = \boldsymbol{I}_{R_2}$, which leads to the orthonormality of $\boldsymbol{\beta}(t)$ in (2.8). Ignoring the approximation error, Model (2.7) can then be written as

$$\begin{aligned} \boldsymbol{y} &= \{\boldsymbol{S} \times_1 \boldsymbol{A} \,\bar{\times}_2\, \boldsymbol{B}^\intercal \boldsymbol{b}(t) \times_3 \boldsymbol{C}\}^\intercal \boldsymbol{x} + \boldsymbol{\epsilon} \\ &= \{\boldsymbol{S} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} \,\bar{\times}_2\, \boldsymbol{b}(t)\}^\intercal \boldsymbol{x} + \boldsymbol{\epsilon}. \end{aligned} \tag{3.9}$$

The above basis expansion enables us to recast the problem of estimating the varying coefficients of reduced model (2.7) as the problem of estimating the parameters $(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, where $\boldsymbol{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$, $\boldsymbol{A} \in \mathbb{R}^{p \times R_1}$ with $\boldsymbol{A}^\intercal \boldsymbol{A} = \boldsymbol{I}_{R_1}$, $\boldsymbol{B} \in \mathbb{R}^{K \times R_2}$ with $\boldsymbol{B}^\intercal \boldsymbol{B} = \boldsymbol{I}_{R_2}$, and $\boldsymbol{C} \in \mathbb{R}^{q \times R_3}$ with $\boldsymbol{C}^\intercal \boldsymbol{C} = \boldsymbol{I}_{R_3}$. Given i.i.d. copies $\{(\boldsymbol{y}_i, \boldsymbol{x}_i, t_i)\}_{i=1}^n$ of $(\boldsymbol{y}, \boldsymbol{x}, t)$, we consider the constrained least squares estimator:

$$\begin{aligned} &\underset{\boldsymbol{S},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}}{\arg\min} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \{\boldsymbol{S} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} \,\bar{\times}_2\, \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{x}_i \right\|_2^2 \\ &\qquad \text{s.t. } \boldsymbol{A}^\intercal \boldsymbol{A} = \boldsymbol{I}_{R_1},\ \boldsymbol{B}^\intercal \boldsymbol{B} = \boldsymbol{I}_{R_2},\ \boldsymbol{C}^\intercal \boldsymbol{C} = \boldsymbol{I}_{R_3}. \end{aligned} \tag{3.10}$$

In (3.9) and (3.10), $\boldsymbol{S} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$ is the Tucker decomposition of a 3rd-order tensor. In particular, letting $\boldsymbol{G} = \boldsymbol{S} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$, we have $\text{rank}_1(\boldsymbol{G}) \le R_1$, $\text{rank}_2(\boldsymbol{G}) \le R_2$, and

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$\text{rank}_3(\boldsymbol{G}) \leq R_3$, where $\text{rank}_d(\cdot)$ denotes the d-rank of a tensor (Kolda and Bader, 2009), $d = 1, 2, 3$. We depict the Tucker decomposition representation of model (3.9) in Figure 3. For more

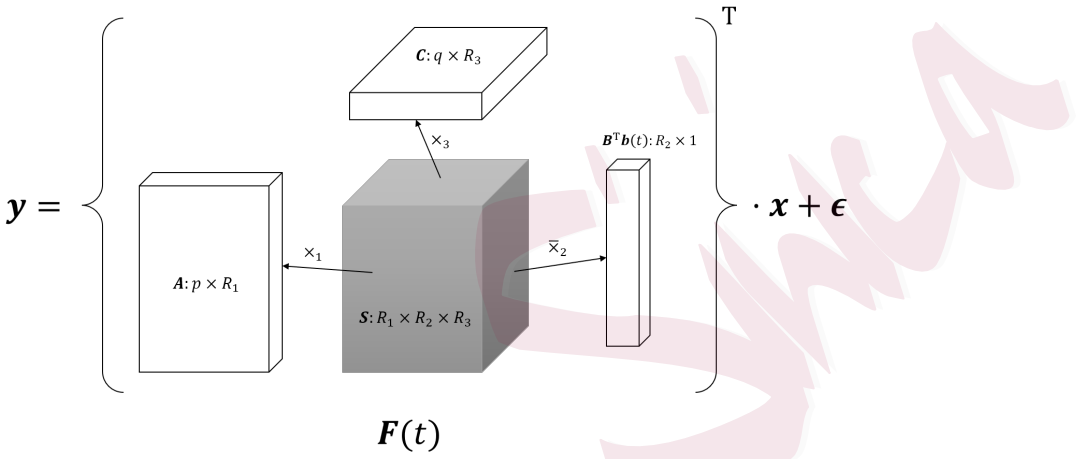

Figure 3: The Tucker decomposition representation of model (3.9).

discussions on the Tucker decomposition and its relationship with other tensor decompositions, such as CANDECOMP/PARAFAC (CP) decomposition (Harshman, 1970) and Tensor-Train decomposition (Oseledets, 2011), we refer the interested readers to Kolda and Bader (2009). Using the form of Tucker decomposition, the least squares problem (3.10) is equivalent to

$$\arg\min_{\boldsymbol{G}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \{\boldsymbol{G} \,\bar{\times}_2\, \boldsymbol{b}(t_i)\}^{\intercal} \boldsymbol{x}_i \right\|_2^2 \quad \text{s.t.} \quad \text{rank}_d(\boldsymbol{G}) \leq R_d,\ d = 1, 2, 3. \tag{3.11}$$

The benefit of using a low-rank structure in tensor regression models against simply flattening the covariate tensor to a matrix or a vector can also been found in Zhou et al. (2013); Li et al. (2018); Ahmed et al. (2020). Note that our problem is different from existing work of Tucker tensor regression (Li et al., 2018) and its generalizations (Lu et al., 2020; Ahmed et al., 2020) in two aspects. First, (3.9) is *not* the proposed model, but merely an approximation of the target *nonparametric* model (2.7). Second, we study multivariate response $\boldsymbol{y}$, whereas Li et al. (2018);

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Lu et al. (2020); Ahmed et al. (2020) all assume the response variable is a scalar.

For a large value of pq, the dimension reduction in terms of low-rank Tucker decomposition may not lead to an accurate estimation for the varying coefficients. In many applications, it is often expected that the responses have similar/related structures and so many important predictors associated with them are shared, and the union of important predictors associated with different responses is of a small size. In other words, we assume that only s ($s < p$ and unknown) predictors are relevant for predicting all the responses. This assumption is shown to be suitable for many real-world applications; see, for example, Wang et al. (2008); Wei et al. (2011); He et al. (2018), among many others. We resort to sparsity-inducing penalization to filter out the irrelevant predictors during estimation. To formulate a suitable penalty function, we use the Tucker decomposition $\boldsymbol{G} = \boldsymbol{S} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$ again and rewrite (3.9) as

$$\boldsymbol{y} = \{\boldsymbol{G} \bar{\times}_2 \boldsymbol{b}(t)\}^\intercal \boldsymbol{x} + \boldsymbol{\epsilon} = \{\boldsymbol{I}_q \otimes \boldsymbol{b}(t)^\intercal\}\boldsymbol{G}_{(1)}^\intercal \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{3.12}$$

where $\boldsymbol{G}_{(1)} \in \mathbb{R}^{p \times qK}$ is the mode-1 matricization (unfolding) of tensor $\boldsymbol{G}$ and $\otimes$ is the Kronecker product of matrices (Kolda and Bader, 2009). Let $\boldsymbol{G}_{(1),j}^\intercal$ denote the j-th row of $\boldsymbol{G}_{(1)}$, $j = 1, \ldots, p$. In light of (3.12), all unknown coefficients associated with the j-th predictor are contained in $\boldsymbol{G}_{(1),j}^\intercal$. Therefore, the j-th predictor becomes irrelevant whenever the coefficient matrix $\boldsymbol{G}_{(1),j}^\intercal = \boldsymbol{0}$. Borrowing the idea from the group Lasso penalization (Yuan and Lin, 2006), we propose the following penalized least squares problem

$$\arg\min_{\boldsymbol{G}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - \{\boldsymbol{G} \bar{\times}_2 \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{x}_i \right\|_2^2 + \sum_{j=1}^{p} \lambda \|\boldsymbol{G}_{(1),j}\|_2, \text{ s.t. } \operatorname{rank}_d(\boldsymbol{G}) \le R_d, \ d = 1, 2, 3, \tag{3.13}$$

where $\|\cdot\|_2$ is the group Lasso penalty and $\lambda \ge 0$ is the penalty parameter. Note that $\boldsymbol{G}_{(1)} = \boldsymbol{A}\boldsymbol{S}_{(1)}(\boldsymbol{C} \otimes \boldsymbol{B})^\intercal$. Let $\boldsymbol{a}_j^\intercal$ be the j-th row of $\boldsymbol{A}$, and then $\boldsymbol{G}_{(1),j}^\intercal = \boldsymbol{a}_j^\intercal \boldsymbol{S}_{(1)}(\boldsymbol{C} \otimes \boldsymbol{B})^\intercal$. Due to

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

the orthonormal conditions of $\boldsymbol{B}$ and $\boldsymbol{C}$, we have $\|\boldsymbol{G}_{(1),j}\|_2 = \|\boldsymbol{a}_j^\intercal \boldsymbol{S}_{(1)}(\boldsymbol{C}\otimes\boldsymbol{B})^\intercal\|_2 = \|\boldsymbol{a}_j^\intercal \boldsymbol{S}_{(1)}\|_2$. Therefore, (3.13) is equivalent to

$$\begin{aligned}\underset{\boldsymbol{S},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}}{\arg\min} \sum_{i=1}^{n} \left\|\boldsymbol{y}_i - \{\boldsymbol{S}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}\,\bar{\times}_2\,\boldsymbol{b}(t_i)\}^\intercal\boldsymbol{x}_i\right\|_2^2 + \sum_{j=1}^{p}\lambda\|\boldsymbol{a}_j^\intercal\boldsymbol{S}_{(1)}\|_2 \\ \text{s.t. } \boldsymbol{A}^\intercal\boldsymbol{A}=\boldsymbol{I}_{R_1},\ \boldsymbol{B}^\intercal\boldsymbol{B}=\boldsymbol{I}_{R_2},\ \boldsymbol{C}^\intercal\boldsymbol{C}=\boldsymbol{I}_{R_3}.\end{aligned} \tag{3.14}$$

Let $(\widehat{\boldsymbol{S}},\widehat{\boldsymbol{A}},\widehat{\boldsymbol{B}},\widehat{\boldsymbol{C}})$ be a solution of (3.14). Correspondingly, a solution of (3.13) can be constructed as $\widehat{\boldsymbol{G}} = \widehat{\boldsymbol{S}}\times_1\widehat{\boldsymbol{A}}\times_2\widehat{\boldsymbol{B}}\times_3\widehat{\boldsymbol{C}}$ (or equivalently $\widehat{\boldsymbol{G}}_{(1)} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{S}}_{(1)}(\widehat{\boldsymbol{C}}\otimes\widehat{\boldsymbol{B}})^\intercal$). The resulting estimated $f_{jl}(t)$ becomes

$$\widehat{f}_{jl}(t) = \sum_{k=1}^{K}\widehat{G}_{jkl}b_k(t), \tag{3.15}$$

where $\widehat{G}_{jkl}$ is the (j,k,l)-th element of $\widehat{\boldsymbol{G}}$. We will provide a theoretical analysis of the proposed estimation in Section 5.

4. Computation

To calculate the estimator, we propose a block updating algorithm to solve the problem (3.14), i.e., updating $\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ alternatively while keeping other components fixed. To facilitate the discussion, we let $\mathcal{L}(\boldsymbol{S},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})$ be the objective function in (3.14) for a given λ, and denote the squared loss and the penalty respectively by

$$H(\boldsymbol{S},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) = \sum_{i=1}^{n}\left\|\boldsymbol{y}_i - \{\boldsymbol{S}\times_1\boldsymbol{A}\times_2\boldsymbol{B}\times_3\boldsymbol{C}\,\bar{\times}_2\,\boldsymbol{b}(t_i)\}^\intercal\boldsymbol{x}_i\right\|_2^2 \text{ and } P(\boldsymbol{S},\boldsymbol{A}) = \sum_{j=1}^{p}\lambda\|\boldsymbol{a}_j^\intercal\boldsymbol{S}_{(1)}\|_2.$$

Denote $\boldsymbol{S}^{(t)}$, $\boldsymbol{A}^{(t)}$, $\boldsymbol{B}^{(t)}$, and $\boldsymbol{C}^{(t)}$ be the t-th iteration ($t\geq 1$) of $\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ in the proposed algorithm respectively. When we update one block with the other blocks fixed, we will use H and/or P with suitable subscripts to simplify the objective functions with respect to the target block.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

For example, when $\boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{C}^{(t)}$ are fixed, we let $H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}) = H(\boldsymbol{S}, \boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{C}^{(t)})$ and $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}) = P(\boldsymbol{S}, \boldsymbol{A}^{(t)})$, respectively, as the functions with respect to $\boldsymbol{S}$. Analogously, we have $H_{\boldsymbol{S}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A})$, $H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$, and $P_{\boldsymbol{S}^{(t+1)}}(\boldsymbol{A})$. The concrete updating details for each block are shown in the following subsections.

4.1 Updating $\boldsymbol{S}$

Using the properties of d-mode product of a tensor (Kolda and Bader, 2009) and vectorization (unfolding of a tensor), we can respectively rewrite $H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S})$ and $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$ as

$$H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}) = \sum_{i=1}^{n} \|\boldsymbol{y}_i - [\boldsymbol{C}^{(t)} \otimes \{\boldsymbol{b}^{\intercal}(t_i)\boldsymbol{B}^{(t)}\} \otimes (\boldsymbol{x}_i^{\intercal}\boldsymbol{A}^{(t)})]\mathrm{vec}\{\boldsymbol{S}_{(1)}\}\|_2^2$$

$$\text{and} \quad P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}) = \lambda \sum_{j=1}^{p} \|(\boldsymbol{a}_j^{(t)})^{\intercal}\boldsymbol{S}_{(1)}\|_2,$$

where $\boldsymbol{S}_{(1)} \in \mathbb{R}^{R_1 \times R_2 R_3}$ is the mode-1 matricization (unfolding) of tensor $\boldsymbol{S}$, $\mathrm{vec}(\cdot)$ is the vectorization operator, and $(\boldsymbol{a}_j^{(t)})^{\intercal}$ is the j-th row of $\boldsymbol{A}^{(t)}$. Thus, updating $\boldsymbol{S}$ is equivalent to obtaining the solution of

$$\min_{\boldsymbol{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}} H_{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{S}) + P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S}). \tag{4.16}$$

Since $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$ is not differentiable, we propose to use a majorization-minimization (MM) algorithm. The acronym can also stand for minorization-maximization if one aims to find the maximum of an objective function; see, for example, Hunter and Lange (2004). MM algorithms are useful extensions of the well-known class of EM algorithms in which the E-step is equivalent to a minorization step. To construct the majorized function for $P_{\boldsymbol{A}^{(t)}}(\boldsymbol{S})$, we extend the MM algorithm of the Lasso penalty (Hunter and Li, 2005) to the group Lasso penalization. Moreover, since (4.16) is an objective function with respect to a tensor, some tensor operations need to

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

be considered and applied to this subproblem. See Section S.1.1 of Supplementary Material for more details, where Algorithm S.1 summarizes the proposed MM algorithm to update $\boldsymbol{S}$.

4.2 Updating $\boldsymbol{A}$

Similar to Section 4.1, we use the properties of vectorization and d-mode product (Kolda and Bader, 2009) to simplify $H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A})$ as

$$H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A}) = \frac{1}{2}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i - \left[\left\{\left[\boldsymbol{C}^{(t)}\otimes\{\boldsymbol{b}^{\intercal}(t_i)\boldsymbol{B}^{(t)}\}\right]\left(\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}\right)^{\intercal}\right\}\otimes\boldsymbol{x}_i^{\intercal}\right]\mathrm{vec}(\boldsymbol{A})\right\|^2.$$

To simplify the updating procedure for $\boldsymbol{A}$, we first remove the orthonormal constraint on $\boldsymbol{A}$ and update $\boldsymbol{A}$ in the Euclidean space. An orthonormalization step will be added in the outer loop to project the updated $\boldsymbol{A}$ back to an orthonormal matrix. The subproblem of $\boldsymbol{A}$ without the orthonormal constraint can then be written as

$$\min_{\boldsymbol{A}}\left\{H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A}) + P_{\boldsymbol{S}^{(t+1)}}(\boldsymbol{A})\right\}, \tag{4.17}$$

where $P_{\boldsymbol{S}^{(t+1)}}(\boldsymbol{A}) = \lambda\sum_{j=1}^{p}\|(\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^{\intercal}\boldsymbol{a}_j\|$. Since there is no analytic solution to (4.17), we propose to use the Alternating Direction Method of Multipliers (ADMM, Gabay and Mercier, 1976). Denote $g(x) = \|x\|$ and introduce the slack variable $\gamma_j \in \mathbb{R}^{R_2R_3}$, $j = 1,\ldots,p$. We rewrite the optimization problem (4.17) as

$$\min_{\boldsymbol{A},\boldsymbol{\Gamma}}\left\{H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A}) + \lambda\sum_{j=1}^{p}g(\boldsymbol{\gamma}_j)\right\}, \quad \text{s.t.} \quad \boldsymbol{\Gamma} = \boldsymbol{A}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}, \tag{4.18}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\ldots,\boldsymbol{\gamma}_p)^{\intercal}$. In (4.18), the constraint is equivalent to $\boldsymbol{\gamma}_j = (\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)})^{\intercal}\boldsymbol{a}_j$, $j = 1,2,\ldots,p$. The corresponding augmented Lagrangian function is

$$\mathcal{L}_\rho(\boldsymbol{A},\boldsymbol{\Gamma};\boldsymbol{\nu}) = H_{\widetilde{\boldsymbol{S}}^{(t+1)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}}(\boldsymbol{A}) + \lambda\sum_{j=1}^{p}g(\boldsymbol{\gamma}_j) + \frac{\rho}{2}\left\|\boldsymbol{A}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)} - \boldsymbol{\Gamma} + \frac{1}{\rho}\boldsymbol{\nu}\right\|_2^2, \tag{4.19}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where $\boldsymbol{\nu} \in \mathbb{R}^{p \times R_2 R_3}$ is the dual variable.

We defer the detailed analysis of (4.19) to Section S.1.2 of Supplementary Material, in which Algorithm S.2 summarizes the proposed ADMM algorithm. Let $\widetilde{\boldsymbol{A}}^{(t+1)}$ denote the output of Algorithm S.2 for $\boldsymbol{A}$. To project the updated $\boldsymbol{A}$ to be an orthonormal matrix, we further let qr.Q($\widetilde{\boldsymbol{A}}^{(t+1)}$) and qr.R($\widetilde{\boldsymbol{A}}^{(t+1)}$) be the Q and R factors of the QR decomposition of $\widetilde{\boldsymbol{A}}^{(t+1)}$, respectively. Here we require the R factor to have positive diagonal elements for QR identifiability. We update $\boldsymbol{A}^{(t+1)}$ as qr.Q($\widetilde{\boldsymbol{A}}^{(t+1)}$), and $\boldsymbol{S}_{(1)}^{(t+1)}$ is then updated as qr.R($\widetilde{\boldsymbol{A}}^{(t+1)}$) $\cdot$ $\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$ (so for $\boldsymbol{S}^{(t+1)}$). Further note that the direct output of Algorithm S.2 does not result in the exact row sparsity of $\widetilde{\boldsymbol{A}}^{(t+1)}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$. To achieve the variable selection in our algorithm, we can further output the slack variable $\boldsymbol{\Gamma}^{(t+1)}$ in Algorithm S.2 as an auxiliary result and replace $\widetilde{\boldsymbol{A}}^{(t+1)}\widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$ by $\boldsymbol{\Gamma}^{(t+1)}$, which indeed fulfills the constraint of the slack variable in (4.18) and the difference of these two terms is close enough. The output of $\boldsymbol{\Gamma}^{(t+1)}$ in Algorithm S.2 remains unchanged when the above orthonormalization step is applied.

4.3 Updating $\boldsymbol{B}$

We let the orthogonal Stiefel manifold be

$$\mathrm{St}(R_2, K) = \{\boldsymbol{B} \in \mathbb{R}^{K \times R_2} : \boldsymbol{B}^\intercal \boldsymbol{B} = \boldsymbol{I}_{R_2}\}. \tag{4.20}$$

Using the properties of the d-mode product of a tensor and a matrix (Kolda and Bader, 2009), we can rewrite $H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B})$ and update $\boldsymbol{B}$ from solving the optimization problem

$$\boldsymbol{B}^{(t+1)} = \underset{\boldsymbol{B} \in \mathrm{St}(R_2,K)}{\arg\min} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - ([\{\boldsymbol{C}^{(t)} \otimes (\boldsymbol{x}_i^\intercal \boldsymbol{A}^{(t+1)})\}\{\boldsymbol{S}_{(2)}^{(t+1)}\}^\intercal] \otimes \boldsymbol{b}^\intercal(t_i))\mathrm{vec}(\boldsymbol{B}) \right\|_2^2, \tag{4.21}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

where $\boldsymbol{S}_{(2)}$ is the mode-2 matricization of tensor $\boldsymbol{S}$. Note that the objective function in (4.21) is a smooth function with respect to $\boldsymbol{B}$ on the Stiefel manifold (4.20), so we can use manifold gradient method (Absil et al., 2009), which is an extension of the gradient descent algorithm in the manifold space. Algorithm S.3 in Section S.1.3 of Supplementary Material specializes our implementation for using the gradient decent algorithm on the Stiefel manifold.

4.4 Updating $\boldsymbol{C}$

Using $\boldsymbol{S}_{(3)}$ as the mode-3 matricization (unfolding) of tensor $\boldsymbol{S}$, we can rewrite (3.9) as

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{S}_{(3)}\{(\boldsymbol{b}^\intercal(t)\boldsymbol{B}) \otimes (\boldsymbol{x}^\intercal\boldsymbol{A})\}^\intercal + \boldsymbol{\epsilon}.$$

Denote $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\intercal \in \mathbb{R}^{n\times q}$ and $\boldsymbol{M}_C^{(t)} = (\boldsymbol{M}_{C,1}^{(t)}, \ldots, \boldsymbol{M}_{C,n}^{(t)})^\intercal \in \mathbb{R}^{n\times R_3}$, where $\boldsymbol{M}_{C,i}^{(t)} = \{\boldsymbol{b}^\intercal(t_i)\boldsymbol{B}^{(t+1)} \otimes (\boldsymbol{x}_i^\intercal\boldsymbol{A}^{(t+1)})\}(\boldsymbol{S}_{(3)}^{(t+1)}) \in \mathbb{R}^{R_3}$, $i = 1, \ldots, n$. We then focus the following subproblem to update $\boldsymbol{C}$

$$\boldsymbol{C}^{(t+1)} = \underset{\boldsymbol{C}^\intercal\boldsymbol{C}=\boldsymbol{I}}{\arg\min} \left\|\boldsymbol{Y} - \boldsymbol{M}_C^{(t)}\boldsymbol{C}^\intercal\right\|_F^2, \tag{4.22}$$

which is known as the Orthonormal Procrustes problem (Gower and Dijksterhuis, 2004), and the solution to this problem is equivalent to finding the nearest orthonormal matrix of $\boldsymbol{Y}^\intercal\boldsymbol{M}_C^{(t)}$. Therefore, write the singular value decomposition of $\boldsymbol{Y}^\intercal\boldsymbol{M}_C^{(t)}$ as

$$\boldsymbol{Y}^\intercal\boldsymbol{M}_C^{(t)} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal, \tag{4.23}$$

where $\boldsymbol{U} \in \mathbb{R}^{q\times R_3}$ and $\boldsymbol{V} \in \mathbb{R}^{R_3\times R_3}$ are orthonormal matrices, and $\boldsymbol{\Sigma} \in \mathbb{R}^{R_3\times R_3}$ is a diagonal matrix with non-negative values in its diagonal. The analytic solution to (4.22) can be obtained as

$$\boldsymbol{C}^{(t+1)} = \boldsymbol{U}\boldsymbol{V}^\intercal. \tag{4.24}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

4.5 Summary and Initializations

We summarize the procedure of the block updating algorithm in Algorithm 1. To achieve a sparse solution, the output of Algorithm 1 is $\widehat{\boldsymbol{G}}_{(1)} = \widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{C}} \otimes \widehat{\boldsymbol{B}})^{\intercal}$. We can then reconstruct $\widehat{\boldsymbol{G}}$ through the estimated $\widehat{\boldsymbol{G}}_{(1)}$ by the inverse of the mode-1 unfolding and obtain the estimator of the varying coefficients using (3.15).

Here are some notable remarks for Algorithm 1. For the subproblems of $\boldsymbol{S}$ and $\boldsymbol{A}$, due to convexity, we can show that the corresponding MM algorithm can generate a sequence converging to the unique minimizer of each subproblem using similar arguments as in Corollary 3.3 of Hunter and Li (2005). We thus use random initializations for $\boldsymbol{S}$ and $\boldsymbol{A}$ at their first iteration of outer loop. Afterwards, we let the outputs of $\boldsymbol{S}$ and $\boldsymbol{A}$ obtained from the preceding iteration of the outer loop be the initialization values, respectively, of the next iteration of the outer loop. As for $\boldsymbol{C}$, the corresponding subproblem for this component can be written as an Orthogonal Procrustes problem which has a closed-form solution, and thus, no initialization is needed for $\boldsymbol{C}$. Finally, the subproblem for $\boldsymbol{B}$ is not convex due to the orthonormal constraint, and the proposed manifold gradient descent algorithm uses only the first-order information on the objective function, which may not guarantee the convergence to a local minimizer (Absil et al., 2009). Therefore, although Algorithm 1 can guarantee to obtain a sequence of decreasing values of the objective function, it is unclear whether this algorithm would guarantee the convergence to a global minimizer. Nevertheless, Absil et al. (2009) also shows that using any sub-sequence of the iterations generated by the manifold gradient descent algorithm can converge to the stationary point of the subproblem. We can thus run Algorithm 1 from multiple initializations of $\boldsymbol{B}$ and return the best result. However, this is computationally expensive. Instead, we propose to use

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

a rough estimator $\boldsymbol{B}^{\text{init}}$ as an initial point for the manifold optimization of $\boldsymbol{B}$. Specifically, at $(t+1)$-th iteration, define

$$\widetilde{\boldsymbol{B}} := \underset{\boldsymbol{B}\in\mathbb{R}^{K\times R_2}}{\arg\min}\, H_{\boldsymbol{S}^{(t+1)},\boldsymbol{A}^{(t+1)},\boldsymbol{C}^{(t)}}(\boldsymbol{B}),$$

which can be easily solved since the objective function is differentiable with respect to $\boldsymbol{B}$ in the Euclidean space. Next, we simply project $\widetilde{\boldsymbol{B}}$ onto the Stiefel manifold and let the projection be the initial point, i.e.,

$$\boldsymbol{B}^{\text{init}} = \mathcal{P}_{\text{St}(R_2,K)}(\widetilde{\boldsymbol{B}}) = \widetilde{\boldsymbol{B}}(\widetilde{\boldsymbol{B}}^{\intercal}\widetilde{\boldsymbol{B}})^{-1/2}.$$

We use the above $\boldsymbol{B}^{\text{init}}$ as the initial value in Algorithm S.3 when we update $\boldsymbol{B}$. Our numerical experiments show that this strategy is not only faster than the use of multiple random initializations but also generates stable iteration sequences.

4.6 Tuning Parameters

Our model has totally six tuning parameters $(m, K, R_1, R_2, R_3, \lambda)$, where m is the order of the spline basis, K is the number of basis, (R_1, R_2, R_3) are the Tucker ranks, and λ is the regularization parameter. We first fix the spline order $m = 4$ (cubic spline) which is commonly used to alleviate the computational burden for estimating nonparametric functions (Ruppert et al., 2003). For the number K of the spline basis functions, there exist many recommended data-driven methods to decide K based on the sample size (see, for example, Huang et al., 2002, 2004; Ruppert et al., 2003, and references therein) for empirical studies. To be computationally simple, we follow the strategy used in Fan et al. (2014) to let $K = [2n^{1/5}]$, where $[\cdot]$ denotes rounding to the nearest integer. The knots of spline basis functions are also data-driven and chosen as

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Algorithm 1: Block Updating Algorithm to Solve (3.14).

Input: Dataset $\{\boldsymbol{y}_i, \boldsymbol{X}_i, t_i\}_{i=1}^n$; Random initial points

$$\boldsymbol{S}^{(0)} \in \mathbb{R}^{R_1 \times R_2 \times R_3}, \boldsymbol{A}^{(0)} \in \mathbb{R}^{p \times R_1}, \boldsymbol{B}^{(0)} \in \mathrm{St}(R_2, K), \boldsymbol{C}^{(0)} \in \mathrm{St}(R_3, q), \text{ and } t = 0.$$

Output: $\widehat{\boldsymbol{G}}_{(1)} = \widehat{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{C}} \otimes \widehat{\boldsymbol{B}})^\intercal$.

repeat

1. Update $\widetilde{\boldsymbol{S}}^{(t+1)}$ using Algorithm S.1.
2. Update $\widetilde{\boldsymbol{A}}^{(t+1)}$ using Algorithm S.2 as well as $\boldsymbol{\Gamma}^{(t+1)}$ for variable selection.
3. After the QR decomposition of $\widetilde{\boldsymbol{A}}^{(t+1)}$, let $\boldsymbol{A}^{(t+1)}$ and $\boldsymbol{S}_{(1)}^{(t+1)}$ be $\mathrm{qr.Q}(\widetilde{\boldsymbol{A}}^{(t+1)})$ and $\mathrm{qr.R}(\widetilde{\boldsymbol{A}}^{(t+1)}) \cdot \widetilde{\boldsymbol{S}}_{(1)}^{(t+1)}$, repectively.
4. Update $\boldsymbol{B}^{(t+1)}$ using manifold gradient optimization method (Algorithm S.3).
5. Update $\boldsymbol{C}^{(t+1)} = \boldsymbol{U}\boldsymbol{V}^\intercal$ as in (4.24) with $\boldsymbol{U}$ and $\boldsymbol{V}$ being defined in (4.23).
6. $t = t + 1$.

until $\mathcal{L}(\boldsymbol{S}^{(t+1)}, \boldsymbol{A}^{(t+1)}, \boldsymbol{B}^{(t+1)}, \boldsymbol{C}^{(t+1)}) - \mathcal{L}(\boldsymbol{S}^{(t)}, \boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{C}^{(t)}) < \epsilon$. Denote $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^{(t+1)}$, $\widehat{\boldsymbol{C}} = \boldsymbol{C}^{(t+1)}$, and $\widehat{\boldsymbol{B}} = \boldsymbol{B}^{(t+1)}$.

equally spaced quantiles. We find this empirical rule works well in all of our experiments. For the choice of R_3, which is corresponding to the dimension reduction associated with the responses, we conduct a singular value decomposition of the response matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$. We then choose R_3 such that the first R_3 dominant singular values, which together account for at least 90% of the sum of all singular values. For (R_1, R_2) and λ, we apply the hold-out method (He et al., 2018; Hannun et al., 2019) in our numerical study for its computational efficiency. More precisely,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

we randomly split our available data into two subsets: a training set with 75% samples and a validation set with 25% samples. We set the validation samples aside, and use Algorithm 1 to fit our proposed method on the training set. The parameters (R_1, R_2) and λ are selected by minimizing the validation error:

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

over the grids of corresponding tuning parameters, where n_{valid} is the size of the validation set, $\widehat{y}_{\text{valid},i}$ is the prediction value of the i-th observation $y_{\text{valid},i}$ in the validation set.

5. Theory

In this section, we establish the oracle inequality for the prediction accuracy of the proposed estimator. For readability, we first show the oracle inequality under a *fixed-design* setting, where the predictors and the exposure variable are fixed. Similarly, we say a setting is *random design* if these variables are randomly distributed. To extend to random design settings, we show that the corresponding assumption on the design (Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$) can be satisfied with high probability (tending to one) when $\boldsymbol{x}$ and t are random under some mild regularity conditions. The result under the fixed-design setting is presented below, while we defer the theoretical result for random design to Section S.4 of Supplementary Material.

Let $\boldsymbol{\Sigma} = \boldsymbol{Z}^\intercal \boldsymbol{Z}/n$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\intercal$, where $\boldsymbol{z}_i = \boldsymbol{x}_i \otimes \boldsymbol{b}(t_i) \in \mathbb{R}^{pK}$. We use $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ to denote the maximum and minimum eigenvalue of a matrix, respectively. Denote by $\boldsymbol{S}_0$, $\boldsymbol{A}_0$, $\boldsymbol{C}_0$, and $\boldsymbol{\beta}_0$ the true values of $\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{C}$, and $\boldsymbol{\beta}$ in (2.7), respectively. Denote by s the number of non-zero rows in $\boldsymbol{A}_0$, which corresponds to the relevant predictors. We also write $\boldsymbol{H}_0 = \boldsymbol{S}_0 \times_1 \boldsymbol{A}_0 \times_3 \boldsymbol{C}_0 \in \mathbb{R}^{p \times R_2 \times q}$, and correspondingly the true coefficient functions are

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

$(f_{0,jl}(t))_{p\times q} = \boldsymbol{F}_0(t) = \boldsymbol{H}_0 \,\bar{\times}_2\, \boldsymbol{\beta}_0(t)$. Let

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^\intercal \in \mathbb{R}^{n\times q} \quad \text{and} \qquad \boldsymbol{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\intercal \in \mathbb{R}^{n\times q}.$$

Now we state a condition that will be required to describe the oracle inequality in our theoretical results.

Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$. *We say the design matrix $\boldsymbol{\Sigma}$ satisfies Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ for an index set $\mathcal{J} \subset \{1, \ldots, p\}$ and a positive number $\delta_{\mathcal{J}}$, if*

$$\mathrm{tr}(\boldsymbol{M}^\intercal \boldsymbol{\Sigma} \boldsymbol{M}) \geq \delta_{\mathcal{J}} \sum_{j\in\mathcal{J}} \|\boldsymbol{M}_j\|_F^2$$

for all $\boldsymbol{M} \in \mathbb{R}^{pK\times q}$ satisfying $2\sum_{j\in\mathcal{J}} \|\boldsymbol{M}_j\|_F \geq \sum_{j\in\mathcal{J}^c} \|\boldsymbol{M}_j\|_F$, where $\boldsymbol{M}_j$ is the collection of rows related to the j-th predictor in $\boldsymbol{M}$ and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ is similar to the one used in Bunea et al. (2012) for reduced rank regression models. In particular, Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ is motivated by the "restricted eigenvalue" (RE) condition introduced in Bickel et al. (2009) for studying the asymptotic properties of high-dimensional linear regression. It implies that the least eigenvalue of relevant predictors is greater than or equal to $\delta_{\mathcal{J}}$ by letting $\boldsymbol{M}_j = \boldsymbol{0}$, $j \in \mathcal{J}^c$. Note that the constant 2 in the inequality $2\sum_{j\in\mathcal{J}} \|\boldsymbol{M}_j\|_F \geq \sum_{j\in\mathcal{J}^c} \|\boldsymbol{M}_j\|_F$ of Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ is merely chosen for neat presentation of the statements, and it can be replaced by any positive constant greater than 1. Lemma S.3 of Supplementary Material shows that when n is at least as large as the magnitude of $|\mathcal{J}|^2q^2K^2 + |\mathcal{J}|^2qK\log p$, Condition $\mathcal{M}(\mathcal{J}, \delta_{\mathcal{J}})$ holds for a constant $\delta_{\mathcal{J}} > 0$ with probability tending to one under some mild conditions of random design.

The following assumptions are needed in our analysis.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Assumption 1. The entries of the noise matrix $\boldsymbol{E}$ are independent and identically distributed Gaussian random variables with mean zero and variance σ^2.

Assumption 2. The columns of the true parameters $\boldsymbol{H}_{0,(2)}$ (mode-2 matricization of $\boldsymbol{H}_0$) has Euclidean norms bounded by a constant.

Assumption 3. The domain of the exposure variable t is $\mathcal{T} = [0, 1]$. The order of the B-spline used in this paper satisfies $\zeta \geq \tau + 1/2$. Let $0 = \xi_1 < \xi_2 < \cdots < \xi_{K-\zeta+2} = 1$ denote the knots of B-spline basis. Furthermore, there exists a positive constant S_1 such that

$$h_n = \max_{k=1,\ldots,K-\zeta+1} |\xi_{k+1} - \xi_k| \asymp K^{-1} \quad \text{and} \quad h_n \Big/ \min_{k=1,\ldots,K-\zeta+1} |\xi_{k+1} - \xi_k| \leq S_1.$$

Assumption 4. The true principal functions $\beta_{0,r_2} \in \mathcal{H}$, $r_2 = 1, \ldots, R_2$. Here $\mathcal{H}$ is the space of functions from $[0, 1]$ to $\mathbb{R}$ satisfying the Hölder condition of order ω, i.e.,

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

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

Assumptions 1–4 are commonly seen in the literature of nonparametric regressions (Huang et al., 2010; He et al., 2018). Specifically, Assumption 1 is used to control the stochastic error. Under Assumptions 3 and 4, it follows from Lemma 5 of Stone (1985) that there exists $\boldsymbol{B}_{0,r_2} = (B_{0,r_2,1}, \ldots, B_{0,r_2,K})^\intercal$ such that for some constant S_2,

$$\left\| \beta_{0,r_2} - \sum_{k=1}^{K} B_{0,r_2,k} b_k \right\|_{\infty} \leq \frac{S_2}{K^{\tau}}, \quad r_2 = 1, \ldots, R_2, \tag{5.25}$$

where $\|\cdot\|_\infty$ is the uniform norm of functions. Let $\boldsymbol{B}_0 = (\boldsymbol{B}_{0,1}, \ldots, \boldsymbol{B}_{0,R_2})^\intercal \in \mathbb{R}^{K \times R_2}$ and

$$\boldsymbol{G}_0 = \boldsymbol{S}_0 \times_1 \boldsymbol{A}_0 \times_2 \boldsymbol{B}_0 \times_3 \boldsymbol{C}_0 \in \mathbb{R}^{p \times K \times q}.$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Note that $\{\boldsymbol{G}_0 \bar{\times}_2 \boldsymbol{b}(t)\}^\intercal \boldsymbol{x}$ is only an approximation to the true regression function, due to the nonparametric nature of the multivariate varying-coefficient model. Using the matricization operator of a tensor (Kolda and Bader, 2009), it can be shown that

$$\{\boldsymbol{G}_0 \bar{\times}_2 \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{x}_i = \boldsymbol{G}_{0,(3)} \boldsymbol{z}_i, \quad i = 1, \ldots, n, \tag{5.26}$$

where $\boldsymbol{G}_{0,(3)}$ is the mode-3 matricization of $\boldsymbol{G}_0$. By (5.25), (5.26), and Assumption 2, the approximation error over n observations, $\boldsymbol{R} := \boldsymbol{Y} - \boldsymbol{E} - \boldsymbol{Z}\boldsymbol{G}_{0,(3)}^\intercal$, satisfies

$$\|\boldsymbol{R}\|_F^2 = \|\boldsymbol{Y} - \boldsymbol{E} - \boldsymbol{Z}\boldsymbol{G}_{0,(3)}^\intercal\|_F^2 \leq S_3 \frac{nsq}{K^{2\tau}}, \tag{5.27}$$

for some positive constant S_3, where we recall s is the number of relevant predictors.

Further, for any $\boldsymbol{G} \in \mathbb{R}^{p \times K \times q}$ with rank restrictions $\text{rank}_d(\boldsymbol{G}) \leq R_d$, $d = 1, 2, 3$, we write

$$\Delta_{\boldsymbol{G}} = \left\{ \sum_{i=1}^n \|\{\boldsymbol{G} \bar{\times}_2 \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{x}_i - \{\boldsymbol{G}_0 \bar{\times}_2 \boldsymbol{b}(t_i)\}^\intercal \boldsymbol{x}_i\|^2 \right\}^{1/2} \tag{5.28}$$

as the discrepancy between $\boldsymbol{G}$ and $\boldsymbol{G}_0$ in terms of prediction. Similarly, we write

$$\Delta_{\boldsymbol{F}} = \left\{ \sum_{i=1}^n \|\boldsymbol{F}(t_i)^\intercal \boldsymbol{x}_i - \boldsymbol{F}_0(t_i)^\intercal \boldsymbol{x}_i\|^2 \right\}^{1/2} \tag{5.29}$$

as the discrepancy between the coefficient functions $\boldsymbol{F}(\cdot)$ with $\boldsymbol{F}(\cdot) = \boldsymbol{G} \bar{\times}_2 \boldsymbol{b}(\cdot)$ and $\boldsymbol{F}_0(\cdot)$. The following Theorem 1 shows the prediction accuracy for a solution $\widehat{\boldsymbol{G}}$ of (3.13), and its proof is deferred to Section S.2 of Supplementary Material.

Theorem 1. *Let $\mathcal{J}(\boldsymbol{G})$ be the index set of nonzero rows of $\boldsymbol{G}_{(1)}$, the mode-1 matricization of $\boldsymbol{G}$ with $\text{rank}_d(\boldsymbol{G}) \leq R_d$, $d = 1, 2, 3$, and denote $R = \min(R_1 R_2, R_3)$. Suppose Assumptions 1–4 hold. Taking*

$$\lambda^2 = S_4 R_3 R n \lambda_{\max}(\boldsymbol{\Sigma}) K \sigma^2 \{1 + \log(p)\} \tag{5.30}$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

for some constant $S_4 > 0$*, we then have*

$$\Delta^2_{\widehat{\boldsymbol{G}}} \leq S_5\Delta^2_{\boldsymbol{G}} + S_6 qR\sigma^2 + S_7\frac{R_3RK|\mathcal{J}(\boldsymbol{G})|\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G})}} + S_8\frac{nsq}{K^{2\tau}} \tag{5.31}$$

with probability at least

$$1 - \frac{8\exp(-q/2)}{3K\log(p)}, \tag{5.32}$$

provided $\boldsymbol{\Sigma}$ *satisfies Condition* $\mathcal{M}(\mathcal{J}(\boldsymbol{G}), \delta_{\mathcal{J}(\boldsymbol{G})})$*, where* $S_5, \ldots, S_8$ *are positive constants.*

Theorem 1 shows the finite-sample oracle inequality for the prediction error between the proposed estimator and its oracle spline approximation. Since the proposed Algorithm 1 cannot guarantee the generated sequence converges to a global minimum of the optimization problem, we remark that there exists a gap between the oracle inequality for the global optimizer and the practical output from the proposed block updating algorithm.

For the coefficient functions, we correspondingly denote $\widehat{\boldsymbol{F}}(t) = \widehat{\boldsymbol{G}} \;\bar{\times}_2\, \boldsymbol{b}(t)$, where $\widehat{\boldsymbol{G}}$ is a solution to (3.13). Theorem 1 can then be generalized to the prediction error for $\widehat{\boldsymbol{F}}(t)$ in terms of (5.29) as the following corollary. The proof of Corollary 1 is deferred to Section S.3 of Supplementary Material.

Corollary 1. *We have*

$$\Delta^2_{\widehat{\boldsymbol{F}}} \leq 2S_5\Delta^2_{\boldsymbol{G}} + 2S_6 qR\sigma^2 + 2S_7\frac{R_3RK|\mathcal{J}(\boldsymbol{G})|\lambda_{\max}(\boldsymbol{\Sigma})\sigma^2\log(p)}{\delta_{\mathcal{J}(\boldsymbol{G})}} + (2S_8 + 2S_3)\frac{nsq}{K^{2\tau}}$$

with probability at least (5.32) *under the same conditions of Theorem 1.*

One direct application of Theorem 1 is to obtain the rate of convergence for the prediction accuracy of the proposed estimator. We can also show that the relevant predictors can be identified with probability tending to one. In the following, let $\|f_{0,jl}\|_2$ be the L_2 norm of $f_{0,jl}$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

under the Lebesgue measure and $\widehat{f}_{jl}$ be the estimated coefficient function of $f_{0,jl}$ from (3.15). The proof of Corollary 2 is deferred to Section S.3 of Supplementary Material.

Corollary 2. *Suppose Assumptions 1–4 hold and* $\boldsymbol{\Sigma}$ *satisfies Condition* $\mathcal{M}(\mathcal{J}(\boldsymbol{G}_0), \delta_{\mathcal{J}(\boldsymbol{G}_0)})$. *If we let*

$$K \asymp \left\{ \frac{n\delta_{\mathcal{J}(\boldsymbol{G}_0)} q}{R_3 R \lambda_{\max}(\boldsymbol{\Sigma}) \log(p)} \right\}^{1/(2\tau+1)},$$

and λ^2 *as in* (5.30), *then the prediction error* $\Delta^2_{\widehat{\boldsymbol{F}}}/n$ *of estimated coefficient functions* $\widehat{\boldsymbol{F}}$ *satisfies*

$$\Delta^2_{\widehat{\boldsymbol{F}}}/n = O_p\left(\frac{qR}{n} + \left\{ \frac{R_3 R \lambda_{\max}(\boldsymbol{\Sigma}) \log(p)}{n\delta_{\mathcal{J}(\boldsymbol{G}_0)}} \right\}^{2\tau/(2\tau+1)} s q^{1/(2\tau+1)} \right). \tag{5.33}$$

Further, if

$$\begin{aligned} &\frac{q^{(4\tau+1)/(2\tau+2)} \delta^{-1/(2\tau+2)}_{\mathcal{J}(\boldsymbol{G}_0)} R \{ R_3 \lambda_{\max}(\boldsymbol{\Sigma}) \log(p) \}^{-\tau/(\tau+1)}}{n} \\ &\quad + \frac{s^{(2\tau+1)/\tau} q^{1/(2\tau)} \delta^{-(4\tau+1)/(2\tau)}_{\mathcal{J}(\boldsymbol{G}_0)} R_3 R \lambda_{\max}(\boldsymbol{\Sigma}) \log(p)}{n} \to 0 \end{aligned} \tag{5.34}$$

as $n \to \infty$ *and* $\sum_{l=1}^{q} \|f_{0,jl}\|_2^2 \geq S_9$ *for some constant* $S_9 > 0$, $\forall j \in \mathcal{J}(\boldsymbol{G}_0)$, *we then have*

$$\mathbb{P}\{ \widehat{\boldsymbol{F}}_j(t) \neq \boldsymbol{0},\ j \in \mathcal{J}(\boldsymbol{G}_0) \} \to 1 \text{ as } n \to \infty,$$

where $\widehat{\boldsymbol{F}}_j^{\intercal}(t) = (\widehat{f}_{j1}, \ldots, \widehat{f}_{jq})$ *is the* j*-th row* $\widehat{\boldsymbol{F}}$.

As we presented in Section 2, Models (2.3) (Lian and Ma, 2013) and (2.4) (He et al., 2018) can be regarded as special cases of our proposed all-mode reduction method. The derived rate of convergence in (5.33) include those in He et al. (2018) and Lian and Ma (2013) as special cases with an extra $\log p$ term due to the use of different penalization method. Condition (5.34) for the variable selection consistency indicates that the sample size n should be large enough compared with the numbers of relevant predictors s and responses q. A simple and sufficient condition for (5.34) to hold is that n should be larger than the magnitude of $q^2 s^4 R_3 R \lambda_{\max}(\boldsymbol{\Sigma}) \log(p) \delta^{-2}_{\mathcal{J}(\boldsymbol{G}_0)}$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

6. Experiments

6.1 Synthetic Data

We conduct a simulation study to evaluate the performance of the proposed model. The data were simulated from the model:

$$y_{il} = \sum_{j=1}^{p} f_{jl}(t_i)x_{il} + \varepsilon_{il}, \quad i = 1, \ldots, n;\ l = 1, \ldots, q,$$

where $\{\varepsilon_{il}\}$ are i.i.d. random variables with normal distribution $\mathcal{N}(0, \sigma^2)$. We set $x_{i1} = 1$ as the intercept for all i, and the remaining $p-1$ predictors were generated from a multivariate Gaussian distribution with mean zero and covariance $\mathrm{Cov}(x_{ij_1}, x_{ij_2}) = \rho^{|j_1 - j_2|}$, $1 \leqslant j_1, j_2 \leqslant p-1$. The exposure variable t_i was generated from the uniform distribution on $[0, 1]$, $i = 1, \ldots, n$. $\{f_{jl}\}$ were generated according to the all-mode reduction model as in (2.7). In particular, the elements of $\boldsymbol{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ and $\boldsymbol{C} \in \mathbb{R}^{q \times R_3}$ were i.i.d. $\mathcal{N}(0, 1)$ random variables. We let the first s predictors, including the intercept, be the truly relevant predictor variables, and the rest $p - s$ predictors have no effect on the responses $\{y_{il}\}$. Therefore, we generated the entries of the first s rows of $\boldsymbol{A} \in \mathbb{R}^{p \times R_1}$ independently from $\mathcal{N}(0, 1)$ and the remaining rows were set as zero.

We set $R_1 = R_2 = R_3 = 2$, $p = 51$ or 201, $s = 11$, $q = 15$, and $\rho = 0.3$. As for σ^2, it was chosen according to the signal-to-noise ratio (SNR), $\mathrm{trace}\{\mathrm{Var}(\sum_{j=1}^{p} f_{jl}(t_i)x_{il})\}/q\sigma^2$. More specifically, two SNRs, 20 and 2, were investigated in our simulation study. The normalized principal functions were specified as $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t))^\intercal = (\sqrt{2}\cos(\pi t), \sqrt{2}\sin(2\pi t))^\intercal$ on the domain $t \in [0, 1]$, which satisfy $\int \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^\intercal \, \mathrm{d}t = \boldsymbol{I}_2$, a 2×2 identity matrix. Two sample sizes, 200 and 400 were considered. For each scenario, 50 replicates were generated.

To fit our model, all the tuning parameters of the proposed method were selected as discussed

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

in Section 4.6. We referred to our proposed method as the all-mode reduction in the following discussion.

We compared the all-mode reduction with four different methods: the mode-3 reduction model (Lian and Ma, 2013), the mode-2 reduction model (He et al., 2018), the full model, and the linear model. Here, the full model refers to (2.2) with the group Lasso method (Yuan and Lin, 2006) to select the relevant predictors. We can set $R_1 = p$, $R_2 = K$, and $R_3 = q$ in our model and use Algorithm S.1 of Supplementary Material to solve its estimator. In the linear model, the regression coefficients are assumed to be constants and the group Lasso method is also employed. Both the full model and the linear model have the tuning parameter λ. To select λ, we used the same validation criteria as our model for the full model and cross-validation for the linear model. The mode-3 reduction model corresponds to dimension reduction in the responses. Therefore, its estimator can be obtained by setting $R_1 = p$ and $R_2 = K$ in our model and iteratively updating $\boldsymbol{S}$ and $\boldsymbol{C}$ using Algorithm S.1 of Supplementary Material and (4.24). The tuning parameters R_3 and λ were selected by the hold-out method. As for the mode-2 reduction model, we applied the implementation provided in He et al. (2018), where cross-validation was used to select the tuning parameters R_2 and λ.

In terms of variable selection, we calculated "True Discovery" as the average number of predictors selected by the methods which in fact are relevant ones, and used "False Discovery" to stand for the average number of predictors selected by the methods which in fact are irrelevant ones. The variable selection performance of the competing methods is summarized in Tables 1 and 2 for sample sizes $n = 250$ and $n = 400$, respectively, together with the performance of the rank selection $\widehat{R}_1$, $\widehat{R}_2$, and $\widehat{R}_3$ for corresponding methods. Note that the reported selected ranks

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

are the average values of 50 replicates. Tables 1 and 2 show that the proposed all-mode reduction model can identify all the non-zero varying coefficient functions with fewest false discovery among competing methods. Though the full and linear models have high accuracy on identifying relevant predictors, their poor performance on false discovery indicate that these methods falsely include many irrelevant predictors in their estimators. The mode-2 and mode-3 reduction methods have similar performance that they sometimes cannot correctly identify the true non-zero varying coefficients, especially when SNR is relatively small. As for the rank selections, it's show that the third rank can be correctly selected as $\widehat{R}_3 = 2$ by our proposed model. For the first and second ranks, we find that more than 75% for the case SNR=20 and 60% for the case SNR=2 are selected as 2 (the true value) among 50 replicates. On average, the proposed all-mode method may tend to select $\widehat{R}_1$ and $\widehat{R}_2$ slightly larger than their true values.

To evaluate the estimation accuracy, we calculated the average integrated squared error (AISE) of various methods as

$$\text{AISE} = \frac{1}{q}\sum_{j=1}^{p}\sum_{l=1}^{q}\int_0^1 \{\widehat{f}_{jl}(t) - f_{jl}(t)\}^2 \, \mathrm{d}t,$$

where $\widehat{f}_{jl}(t)$ denotes a generic estimator of $f_{jl}(t)$ using different methods. The above integrals were computed by Monte-Carlo method. Table 3 reports the AISEs of competing methods with corresponding standard errors. For benchmark, we also added the oracle estimator which only included the true relevant predictors in its model. In the oracle setting, the true relevant predictors were assumed to be known, and therefore we did not include the penalization in the objective function, which enabled us to use least squares method to estimate $\boldsymbol{S}$ and $\boldsymbol{A}$. We used the same framework of block updating Algorithm 1 to compute the oracle estimator. The boxplot of AISEs for different methods with sample size $n = 400$ was depicted at Figure 4. We

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

			$\widehat{R}_1$	$\widehat{R}_2$	$\widehat{R}_3$	True Discovery	False Discovery
$p = 51$	SNR=20	All-mode Reduction	2.34	2.14	2.00	11.00 (0.00)	3.26 (0.28)
		Mode-3 Reduction	-	-	2.00	10.86 (0.55)	10.45 (0.83)
		Mode-2 Reduction	-	2.19	-	11.00 (0.00)	11.39 (0.70)
		Full Model	-	-	-	11.00 (0.00)	13.48 (0.93)
		Linear Model	-	-	-	11.00 (0.00)	19.07 (1.19)
	SNR=2	All-mode Reduction	2.64	2.34	2.00	11.00 (0.00)	3.65 (0.34)
		Mode-3 Reduction	-	-	2.00	10.34 (0.95)	14.43 (1.14)
		Mode-2 Reduction	-	2.25	-	11.00 (0.00)	18.75 (1.40)
		Full Model	-	-	-	11.00 (0.00)	21.31 (1.54)
		Linear Model	-	-	-	11.00 (0.00)	24.87 (1.60)
$p = 201$	SNR=20	All-mode Reduction	2.67	2.58	2.00	11.00 (0.00)	12.08 (0.59)
		Mode-3 Reduction	-	-	2.00	10.96 (0.54)	19.40 (1.15)
		Mode-2 Reduction	-	2.41	-	11.00 (0.00)	25.86 (1.46)
		Full Model	-	-	-	11.00 (0.00)	28.47 (1.96)
		Linear Model	-	-	-	11.00 (0.00)	31.51 (2.57)
	SNR=2	All-mode Reduction	2.57	2.60	2.00	10.44 (0.39)	15.42 (1.32)
		Mode-3 Reduction	-	-	2.00	9.96 (1.21)	19.87 (1.47)
		Mode-2 Reduction	-	2.84	-	9.83 (0.81)	23.17 (1.74)
		Full Model	-	-	-	11.00 (0.00)	36.48 (2.18)
		Linear Model	-	-	-	11.00 (0.00)	44.87 (2.45)

Table 1: Dimension reduction and variable selection results for group Lasso penalized estimators for $n = 200$. The numbers in the brackets are the standard errors based on 50 replicates.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

			$\widehat{R}_1$	$\widehat{R}_2$	$\widehat{R}_3$	True Discovery	False Discovery
$p = 51$	SNR=20	All-mode Reduction	2.31	2.15	2.00	11.00 (0.00)	2.42 (0.28)
		Mode-3 Reduction	-	-	2.00	11.00 (0.00)	8.64 (0.61)
		Mode-2 Reduction	-	2.20	-	11.00 (0.00)	8.78 (0.66)
		Full Model	-	-	-	11.00 (0.00)	11.52 (0.84)
		Linear Model	-	-	-	11.00 (0.00)	17.87 (1.15)
	SNR=2	All-mode Reduction	2.39	2.21	2.00	11.00 (0.00)	3.71 (0.34)
		Mode-3 Reduction	-	-	2.00	10.52 (0.78)	10.72 (0.90)
		Mode-2 Reduction	-	2.23	-	11.00 (0.00)	12.38 (0.92)
		Full Model	-	-	-	11.00 (0.00)	16.86 (1.01)
		Linear Model	-	-	-	11.00 (0.00)	21.87 (1.45)
$p = 201$	SNR=20	All-mode Reduction	2.45	2.50	2.00	11.00 (0.00)	10.71 (0.58)
		Mode-3 Reduction	-	-	2.00	11.00 (0.00)	18.85 (1.02)
		Mode-2 Reduction	-	2.38	-	11.00 (0.00)	20.32 (1.38)
		Full Model	-	-	-	11.00 (0.00)	23.10 (1.82)
		Linear Model	-	-	-	11.00 (0.00)	30.51 (2.47)
	SNR=2	All-mode Reduction	2.45	2.70	2.00	10.50 (0.24)	11.83 (1.00)
		Mode-3 Reduction	-	-	2.00	10.12 (1.11)	17.57 (1.16)
		Mode-2 Reduction	-	2.49	-	10.33 (0.93)	19.48 (1.36)
		Full Model	-	-	-	11.00 (0.00)	31.59 (2.28)
		Linear Model	-	-	-	11.00 (0.00)	43.07 (1.82)

Table 2: Similar to Table 1 but for $n = 400$.

can conclude from Table 3 and Figure 4 that all-mode reduction model outperforms the non-oracle estimators with smallest AISE. For example, when the sample size $n = 400$, the all-mode

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

reduction method reduces AISE by 48%–94% compared with the mode-3 reduction and by 78%–98% compared with the mode-2 reduction. The performance of the all-mode reduction method is improved when the sample size increases, which is consistent with our theoretical investigation. Among the competing methods, the full model and the linear model show the worst performance.

n	p	SNR	Oracle	All-mode Reduction	Mode-3 Reduction	Mode-2 Reduction	Full Model	Linear Model
200	51	20	0.007	0.011	0.237	0.782	3.459	6.761
			(0.002)	(0.003)	(0.019)	(0.082)	(0.339)	(0.674)
		2	0.031	0.085	0.314	1.484	6.348	10.197
			(0.004)	(0.007)	(0.015)	(0.104)	(0.454)	(0.568)
	201	20	0.008	0.223	0.496	0.794	4.327	15.192
			(0.004)	(0.051)	(0.052)	(0.084)	(0.453)	(0.961)
		2	0.042	0.293	0.615	2.940	10.361	20.387
			(0.006)	(0.009)	(0.063)	(0.281)	(0.972)	(1.623)
400	51	20	0.004	0.010	0.164	0.501	2.240	4.933
			(0.001)	(0.002)	(0.011)	(0.042)	(0.268)	(0.469)
		2	0.018	0.022	0.281	0.841	4.418	8.910
			(0.002)	(0.002)	(0.016)	(0.065)	(0.399)	(0.457)
	201	20	0.005	0.101	0.403	0.679	4.229	14.584
			(0.001)	(0.007)	(0.042)	(0.049)	(0.532)	(0.567)
		2	0.022	0.286	0.549	1.306	9.061	17.178
			(0.002)	(0.009)	(0.065)	(0.118)	(0.895)	(1.340)

Table 3: AISEs for competing methods. The numbers in brackets are the standard errors based on 50 replicates.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

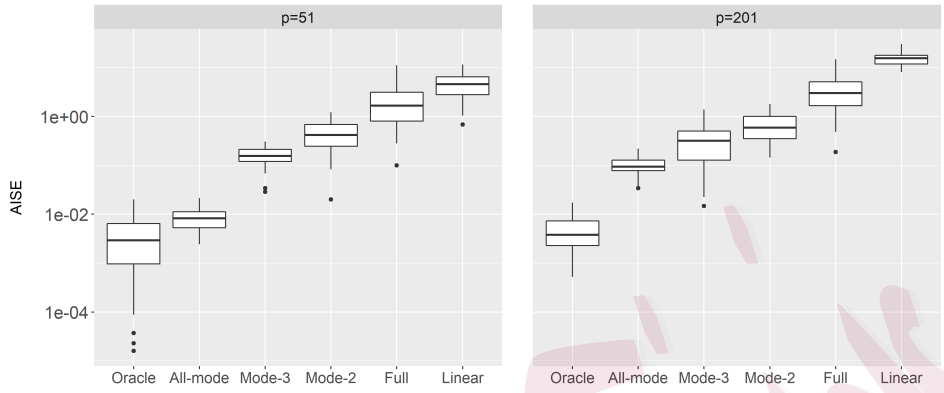

Figure 4: The boxplot of AISEs for different models when $n = 400$ with signal-to-noise ratio being 20. The left and right panels represent the AISEs for $p = 51$ and for $p = 201$, respectively. The y-axis is measured in logarithmic scale.

6.2 Real Data

We further illustrate the proposed method on the dataset of Framingham Heart Study (FHS; Dawber et al. 1951), which aims to identify the common factors that leads to cardiovascular diseases. We used a subset of the data collected from 325 patients, that contains measurements on 15 phenotypes in addition to the Single Nucleotide Polymorohism (SNP) information. All the variables were standardized with mean 0 and variance 1. After matching the SNP data with phenotypes and deleting observations with missing values and outliers, there were 258 patients in our analysis. We preselected 6 phenotypes that we were interested in. They are *height*, *bi-deltoid girth*, *right arm girth-upper third*, *waist girth*, *hip girth*, and *thigh girth*. The exposure variable is set to be the *weight*. We followed the screening procedure in Fan et al. (2014) to select 200 SNPs as predictors (the intercept was also included in the model). To fit our proposed method,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

all the tuning parameters were selected as discussed in Section 4.6. Specifically, the dataset was randomly split into three subsets, i.e., a training set, a validation set, and a test set, of size 150, 50, and 58, respectively. The training and validation sets were used to determine (R_1, R_2) and λ, and the test set was for evaluating the out-sample prediction performance. The recommended rule $K = [2n^{1/5}]$ for the number of basis functions leads to $K = 6$. To evaluate the performance, the corresponding prediction error was defined as

$$\text{Prediction Error} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \|\boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i\|_2^2,$$

where $\boldsymbol{y}_i$'s were the observed responses in the test set, $\widehat{\boldsymbol{y}}_i = \{\widehat{\boldsymbol{G}} \,\bar{\times}_2\, \boldsymbol{b}(t)\}^\intercal \boldsymbol{x}_i$ with $\widehat{\boldsymbol{G}}$ from various methods and $\boldsymbol{x}_i$ the corresponding predictors, and n_{test} was the size of the test set. We compared the proposed model, the all-mode reduction, with 4 non-oracle alternatives in Section 6.1. Furthermore, we additionally implement the elementwise-sparsity method on the full model to fit the dataset of Framingham Heart Study. Here, a full model with elementwise-sparsity method can be achieved by using the group Lasso penalization (Yuan and Lin, 2006) on each coefficient function in (2.2) to select the relevant predictors for different response variables. The performance of each method was evaluated based on 50 random splittings of training, validation, and test sets.

Table 4 records the average prediction error of competing methods on the test data and the performance of the dimension reduction. We observe in Table 4 that the full model with row-sparsity method outperforms that of elementwise-sparsity method, which implies that the dataset of Framingham Heart Study may be better fitted under the row-sparsity method rather than the elementwise-sparsity method. It also shows that the proposed all-mode reduction model has the highest prediction accuracy, and also reveals significant dimensionality reduction on each mode.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

This result is consistent with that of synthetic data. To investigate biological interpretation on the identified SNPs, we input the submitted $ss\sharp$ of the identified SNPs to NCBI database (Sherry et al., 2001) to retrieve the reference $rs\sharp$ records. The proposed all-mode reduction method identified 30 SNPs by combing the variable selection results of 50 random splits and some of them have been scientifically confirmed. For example, the reference SNP $rs4896044$ is found to be associated with hypertension (Consortium, 2007), and $rs9321440$ has links with multiple heart diseases (Gagliardi, 2011). Meanwhile, the mode-3 reduction method identified 51 SNPs, including all of the 30 SNPs selected by the all-mode reduction method. On the other hand, the mode-2 reduction method identified 47 SNPs. 25 SNPs selected by the all-mode reduction method are among these 47 SNPs, including the scientifically confirmed $rs4896044$ and $rs9321440$.

	Prediction error	$\widehat{R}_1$	$\widehat{R}_2$	$\widehat{R}_3$
All-mode Reduction	0.4542 (0.0071)	2.7	3.1	2.0
Mode-3 Reduction	0.6011 (0.0196)	-	-	2.0
Mode-2 Reduction	0.6385 (0.0357)	-	4.3	-
Full Model (row-sparsity)	1.0181 (0.0417)	-	-	-
Full Model (elementwise-sparsity)	1.2106 (0.0403)	-	-	-
Linear Model	1.2578 (0.0488)	-	-	-

Table 4: Prediction error of the test data. The numbers in brackets are the standard errors based on 50 random splitting.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

6.3 Additional Numerical Results

To further demonstrate the utility of the proposed all-mode reduction method, we conducted additional numerical experiments and present the results in Section S.5 of Supplementary Material. More precisely, we extended our simulation settings to larger numbers of response variables q, and plot the trend of the performance of the proposed method when q increases in Section S.5.1 of Supplementary Material. In Section S.5.2 of Supplementary Material, we depict the fitted coefficient functions of the biologically confirmed SNP rs9321440 based on 50 replicates of random splitting. It shows that rs9321440 may have different effects on the phenotypes of *height*, *bi-deltoid girth*, *right arm girth-upper third*, *hip girth*, and *thigh girth* given distinct body weights. As for the phenotype of *waist girth*, the effect of this SNP may not vary with body weights significantly. We refer the interested readers to Section S.5 of Supplementary Material for details.

7. Discussion

In this paper, a dimension-reduction method based on Tucker decomposition of a 3rd-order tensor is proposed to estimate the varying coefficients of multivariate varying-coefficient models under high-dimensional settings. The proposed model unifies dimensionality reductions in three aspects: relevant predictors, coefficient functions, and responses. A sparsity-inducing penalization is also integrated into the estimation due to sparsity consideration. The oracle inequality for the prediction risk of the proposed estimator is derived under the settings of fixed and random designs. We have used both simulated and real data sets to evaluate the empirical performance of the proposed model with some comparison methods, and the results illustrate the superior

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

performance of our method.

One difficulty of applying the proposed method is the need to tune the ranks of Tucker decomposition, which may become computationally expensive when the dimension is extremely high. Developing an efficient way to tune the ranks requires further investigation. Furthermore, in some applications, the relationships between responses can be determined by some external covariates, such as spatial locations. The external covariates, sometimes, provide extra information for measuring the similarity between responses and thus induces a (weighted) graphical structure among tasks. Thus, another interesting future research topic is extending the proposed model to the problem of graph regularized multi-task learning. Finally, incorporating the elementwise-sparsity method with our proposed all-mode reduction model may be useful in other real applications. We also leave this approach as a future research topic.

Acknowledgment

The authors thank the editor, the associate editor, and the anonymous reviewers for their comments that helped significantly improve this work. This research was supported by Public Computing Cloud, Renmin University of China. The research of Kejun He was partially supported by National Natural Science Foundation of China (No.11801560).

References

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

Ahmed, T., H. Raja, and W. U. Bajwa (2020). Tensor regression using low-rank and sparse tucker decompositions. *SIAM*

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Journal on Mathematics of Data Science 2(4), 944–966.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

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

Bunea, F., Y. She, and M. H. Wegkamp (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics 40*(5), 2359–2388.

Consortium, T. W. T. C. C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*(7145), 661–678.

Dawber, T. R., G. F. Meadors, and F. E. Moore Jr (1951). Epidemiological approaches to heart disease: The framingham study. *American Journal of Public Health and the Nations Health 41*(3), 279–286.

Edelman, A., T. A. Arias, and S. T. Smith (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications 20*(2), 303–353.

Fan, J., Y. Ma, and W. Dai (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association 109*(507), 1270–1284.

Gabay, D. and B. Mercier (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications 2*(1), 17–40.

Gagliardi, L. (2011). *Regulation of cortisol secretion in humans: Relation to vasopressin action at the adrenals in macronodular and micronodular adrenocortical tumours; and well-being in Addison's Disease.* Ph. D. thesis.

Gower, J. C. and G. B. Dijksterhuis (2004). *Procrustes problems.* Oxford, UK: Oxford University Press.

Hannun, A. Y., P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng (2019). Cardiologist-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine 25*(1), 65–69.

Harshman, R. (1970). Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics 16*, 1–84.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(4), 757–779.

He, K., H. Lian, S. Ma, and J. Z. Huang (2018). Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. *Journal of the American Statistical Association 113*(522), 746–754.

He, L., F. Wang, K. Chen, W. Xu, and J. Zhou (2018). Boosted sparse and low-rank tensor regression. In *Advances in Neural Information Processing Systems*, Volume 31, pp. 1009–1018.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

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

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*(1), 111–128.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica 14*(3), 763–788.

Hunter, D. R. and K. Lange (2004). A tutorial on mm algorithms. *The American Statistician 58*(1), 30–37.

Hunter, D. R. and R. Li (2005). Variable selection using mm algorithms. *The Annals of Statistics 33*(4), 1617–1642.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Jiang, Q., H. Wang, Y. Xia, and G. Jiang (2013). On a principal varying coefficient model. *Journal of the American Statistical Association 108*(501), 228–236.

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

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

Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica 22*(4), 1563–1588.

Lian, H. and S. Ma (2013). Reduced-rank regression in sparse multivariate varying-coefficient models with high-dimensional covariates. *arXiv preprint arXiv:1309.6058*.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

Lu, W., Z. Zhu, and H. Lian (2020). High-dimensional quantile tensor regression. *Journal of Machine Learning Research 21*(250), 1–31.

Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing 33*(5), 2295–2317.

Park, B. U., E. Mammen, Y. K. Lee, and E. R. Lee (2015). Varying coefficient regression models: A review and new developments. *International Statistical Review 83*(1), 36–64.

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

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin (2001). Dbsnp: The ncbi database of genetic variation. *Nucleic Acids Research 29*(1), 308–311.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics 3*(2), 689–705.

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

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wei, F., J. Huang, and H. Li (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica 21*(4), 1515–1540.

Xue, L. and A. Qu (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research 13*(1), 1973–1998.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

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

Fengyu Zhang and Kejun He (Email: kejunhe@ruc.edu.cn)

Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

Ya Zhou

Department of Information Center, Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China.

Raymond K. W. Wong

Department of Statistics, Texas A&M University, College Station 77843, USA.