Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Dimensionality Reduction and Variable Selection in Multivariate Varying-Coefficient Models with a Large Number of Covariates

Kejun He, Heng Lian, Shujie Ma & Jianhua Z. Huang

To cite this article: Kejun He, Heng Lian, Shujie Ma & Jianhua Z. Huang (2017): Dimensionality Reduction and Variable Selection in Multivariate Varying-Coefficient Models with a Large Number of Covariates, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1285774

To link to this article: http://dx.doi.org/10.1080/01621459.2017.1285774

Dimensionality Reduction and Variable Selection in Multivariate Varying-Coefficient Models with a Large Number of Covariates

Kejun He*, Heng Lian*, Shujie Ma† and Jianhua Z. Huang

Abstract

Motivated by the study of gene and environment interactions, we consider a multivariate response varying-coefficient model with a large number of covariates. The need of nonparametrically estimating a large number of coefficient functions given relatively limited data poses a big challenge for fitting such a model. To overcome the challenge, we develop a method that incorporates three ideas: i. reduce the number of unknown functions to be estimated by using (non-centered) principal components; ii. approximate the unknown functions by polynomial splines; iii. apply sparsity-inducing penalization to select relevant covariates. The three ideas are integrated into a penalized least squares framework. Our asymptotic theory shows that the proposed method can consistently identify relevant covariates and can estimate the corresponding coefficient functions with the same convergence rate as when only the relevant variables are included in the model. We also develop a novel computational algorithm to solve the penalized least squares problem by combining proximal algorithms and optimization over Stiefel manifolds. Our method is illustrated using data from Framingham Heart Study.

Keywords: Multivariate regression; Oracle property; Polynomial splines

Short title: Multivariate Varying-Coefficient Models

Kejun He (Email: mailto:zjhekejun@163.comzjhekejun@163.com) is Assistant Professor, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China. Heng Lian (Email: mailto:henglian@cityu.edu.hkhenglian@cityu.edu.hk) is Associate Professor, Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong. Shujie Ma (Email: mailto:shujie.ma@ucr.edushujie.ma@ucr.edu) is Assistant Professor, Department of Statistics, University of California-Riverside, Riverside, CA 92521. Jianhua Huang (Email: mailto:jianhua@stat.tamu.edujianhua@stat.tamu.edu) is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Ma's research was partially supported by NSF grant DMS 1306972. Huang's work was partially supported by NSF grant DMS 1208952.

*The first two authors contributed equally to this work.

†Corresponding author.

1 Introduction

This work is motivated by developing more flexible statistical models for the study of gene and environment interactions. It has been known over decades that some diseases are linked with genetics factors; for example, study results for hypertension are reported in Kurtz and Spence (1993) and Burton et al. (2007). These genetic effects, however, can be altered under variations of environmental exposures such as stress and dietary factors (Pausova et al., 1999) or BMI (Taylor et al., 2010). Statistical analysis on how the effects of genetics change with the environment is called gene and environment (G×E) interactions (Tabery, 2007).

A conventional method to study the G×E interactions with univariate response is using a linear model. For the i-th subject, $i = 1, \cdots, n$, let Y_i be the value of some phenotype response, $\mathbf{X}_i$ be the $(p+1)$ genetic factors, and T_i be the environmental factor, respectively. The linear regression model with interaction effects can be written as

$$\begin{aligned}\mathbb{E}(Y_i|\mathbf{X}_i, T_i) &= \alpha_0^{(0)} + \alpha_0^{(1)} T_i + \sum\nolimits_{j=1}^{p} \alpha_j^{(0)} X_{ij} + \sum\nolimits_{j=1}^{p} \alpha_j^{(1)} T_i X_{ij} \\ &= \sum\nolimits_{j=0}^{p} (\alpha_j^{(0)} + \alpha_j^{(1)} T_i) X_{ij}, \end{aligned} \tag{1}$$

where $\alpha_j^{(0)}$ and $\alpha_j^{(1)}$ are unknown coefficients.

Although this linear model has a simple form and is convenient to estimate and interpret, it is usually not sufficiently flexible for predicting the phenotype. Especially under the influence of the environmental factor, the strong linearity assumption is easily violated. To see this, Figure 1 shows the estimated mean curve of one trait (which is weight) by cubic splines against hours of sedentary activity per day for the three genotype categories of the SNP ss66101769 from the Framingham Heart Study (Dawber et al., 1951), indicating clear nonlinear interaction effects.

To flexibly model the nonlinear G×E interaction, Ma et al. (2015) proposed to use the generalized varying coefficient model (VCM) where the coefficient functions are specified as nonparametric additive models. As a flexible yet still interpretable extension of the linear model, VCM has been extensively studied in the statistics literature and widely used in practice. Published work on

this subject include Hastie and Tibshirani (1993), Hoover et al. (1998), and Huang et al. (2002), among many others. The VCM in the high-dimensional data settings have been studied in Wei et al. (2011), Lian (2012), Xue and Qu (2012), Fan et al. (2014), and Liu et al. (2014). In the application of VCM to the G×E interactions, Ma et al. (2015) also proposed a method for selecting relevant genes from a large number of candidates.

In reality, however, there are usually multiple phenotypes. For example, in Framingham Heart Study (Dawber et al., 1951), multiple phenotype variables have been collected from some patients. Applying a univariate response VCM to the l-th response, $1 \leq l \leq q$, we have that,

$$Y_{il} = \sum_{j=0}^{p} f_j^{(l)}(T_i)X_{ij} + E_{il} = \mathbf{f}^{(l)}(T_i)^{\mathrm{T}}\mathbf{X}_i + E_{il}, \tag{2}$$

where $\mathbf{f}^{(l)}(t) = \{f_0^{(l)}(t), \ldots, f_p^{(l)}(t)\}^{\mathrm{T}}$, and E_{il} is a mean-zero random error, $1 \leq i \leq n$. To write the model (2) in a matrix form, we further let $\mathbf{Y}_i = \left(Y_{i1}, \ldots, Y_{iq}\right)^{\mathrm{T}}$. Combining all responses in one equation, we obtain the multivariate varying coefficient model (MVCM):

$$\mathbf{Y}_i = \mathbf{F}(T_i)\mathbf{X}_i + \mathbf{E}_i, \tag{3}$$

where $\mathbf{F}(t) = [\{\mathbf{f}^{(1)}(t), \ldots, \mathbf{f}^{(q)}(t)\}^{\mathrm{T}}]_{q\times(p+1)}$ is a matrix of varying coefficients and $\mathbf{E}_i = (E_{i1}, \ldots, E_{iq})^{\mathrm{T}}$. We refer to T_i as the index variable, which is chosen to be "sedentary activity hours" in our application of MVCM to the Framingham Heart Study in Section 3. Parameter estimation would be a challenge for the MVCM since we need to estimate totally $q(p+1)$ coefficient functions, especially when we need to deal with the case that the number of responses and the number of covariates are large. In our asymptotic theory, we allow both q and p, the dimensions of the responses and the covariates, to diverge with the sample size n.

To have an interpretable model and sufficiently reduce the number of unknown functions to be estimated, we borrow the idea of (non-centered) functional principal components analysis as follows. We represent all the coefficient functions with r principal component functions $\{\beta_1, \ldots, \beta_r\}$ such that

$$f_j^{(l)}(t) = \sum_{\nu=1}^{r} d_{jl,\nu}\beta_\nu(t), \qquad 1 \leq l \leq q, 0 \leq j \leq p, \tag{4}$$

where $d_{jl,v}$ are the principal component loadings. Here we use the non-centered principal components analysis because the coefficient functions are not an iid sample from a population of functions and so it is not reasonable to assume a mean function. This reduces the problem of estimating $q(p+1)$ unknown functions in the MVCM model (3) to the problem of estimating r unknown principal component functions and the associated principal component loading matrices. The principal component functions are subject to the orthogonality constraint $\int_{\mathcal{T}} \beta_j \beta_l = \delta_{jl}$, with $\mathcal{T}$ being a compact interval that all T_i's take values, and δ_{jl} being the Kronecker δ. Consequently, we obtain the following reduced multivariate varying-coefficient model (reduced MVCM):

$$\mathbf{Y}_i = \{\mathbf{D}^1\beta_1(T_i) + \cdots + \mathbf{D}^r\beta_r(T_i)\}\mathbf{X}_i + \mathbf{E}_i, \tag{5}$$

where $\mathbf{D}^1, \cdots, \mathbf{D}^r$ are $q \times (p+1)$ matrices. The conventional linear regression model with interaction effects (1) can be viewed as a special case of (5). In particular, when the domain of T is $[0, 1]$, let $\beta_1(t) = 1$ and $\beta_2(t) = \sqrt{6}(t - 1/2)$, then (5) becomes

$$\mathbf{Y}_i = \{\mathbf{D}^1\beta_1(T_i) + \mathbf{D}^2\beta_2(T_i)\}\mathbf{X}_i + \mathbf{E}_i = (\boldsymbol{\alpha}^{(0)} + \boldsymbol{\alpha}^{(1)}T_i)\mathbf{X}_i + \mathbf{E}_i$$

for some $q \times (p + 1)$ coefficient matrices $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\alpha}^{(1)}$, recovering the model (1) in the case of multivariate responses.

One could reduce the burden of estimating too many unknown functions in the original MVCM (3) by representing all coefficient functions using a fixed common basis such as B-splines. In order to have enough flexibility, the basis should be rich enough, i.e., have a large enough dimension, denoted as K. The resulting model, referred to as the full MVCM, has the same form as (5) with r replaced by K and $\boldsymbol{\beta}_i$'s interpreted as the fixed basis functions. It is clear that this fixed basis approach can have much more basis coefficient matrices $\mathbf{D}^j$ to estimate than the reduced MVCM. In contrast, our proposed (reduced MVCM) approach usually only needs to estimate a much smaller number of coefficient matrices together with some data-driven basis functions (i.e., the principal components functions).

After reducing the number of unknown functions to a small number of principal components,

we still need to estimate the basis coefficient matrix $\mathbf{D}^j$'s in (5). Accurate estimation of these $q \times (p+1)$-dimensional matrices is difficult for a typical sample size, especially when p is large. To overcome this difficulty, we assume sparsity of these matrices so that only a few covariates are relevant for prediction of the responses. We are able to show that using a sparsity-inducing penalty, the penalized least squares estimator enjoys the nonparametric oracle property, that is, the irrelevant variables can be consistently identified and corresponding coefficient functions can be estimated with the same convergence rate as when only the relevant variables are included in the model. For computation, by rewriting the penalized least squares criterion in an equivalent form using Kronecker product of matrices (see Section 2.1), we are able to single out a low dimensional manifold structure in a high-dimensional vector-valued function space and thereby develop an iterative algorithm that involves novel applications of the proximal algorithms and optimization through Stiefel manifolds

The reduced MVCM is connected with the multivariate regression as follows. We can rewrite (5) as:

$$\mathbf{Y}_i = \mathbf{D}^1\beta_1(T_i)\mathbf{X}_i + \cdots + \mathbf{D}^r\beta_r(T_i)\mathbf{X}_i + \mathbf{E}_i \tag{6}$$

If $\{\beta_1, \ldots, \beta_r\}$ were known, this model would be a multivariate linear regression with $(\mathbf{D}^1, \cdots, \mathbf{D}^r)$ as the coefficient matrix and $\{\beta_1(T_i)\mathbf{X}_i^{\mathrm{T}}, \cdots, \beta_r(T_i)\mathbf{X}_i^{\mathrm{T}}\}^{\mathrm{T}}$ as the covariates respectively. Variable selection for multivariate linear regression using penalization has been studied by Bunea et al. (2012); Chen et al. (2012); Chen and Huang (2012); Ma et al. (2014, 2016), among others. The need for estimating unknown principal component functions in our reduced MVCM distinguishes this paper from those work. Another related work is Jiang et al. (2013), where the same varying-coefficient model as (2) and (3) were studied but with univariate response. The multiple-step procedure proposed in that paper can not be applied directly to deal with large number of covariates.

The rest of this paper is organized as follows. Section 2 describes the proposed method, including the penalized least squares estimation, the computational algorithm and its convergence analysis, the asymptotic properties, and the simulation study. Section 3 applies the proposed meth-

ods on the real data from the Framingham Heart Study. Some concluding remarks are given in Section 4. The technical proofs of the theoretical results are deferred to the Appendices.

2 Method

2.1 Penalized Least Squares

To facilitate parameter estimation of our reduced MVCM (5), we first rewrite the model in a more succinct form using matrix multiplications. Note that for each $\nu \in \{1, \cdots, r\}$, $\mathbf{D}^{\nu}\beta_{\nu}(T_i)\mathbf{X}_i$ is a vector, thus

$$\mathbf{D}^{\nu}\beta_{\nu}(T_i)\mathbf{X}_i = \text{vec}\{\mathbf{D}^{\nu}\beta_{\nu}(T_i)\mathbf{X}_i\} = (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\text{vec}(\mathbf{D}^{\nu})\beta_{\nu}(T_i),$$

where $\otimes$ and $\text{vec}(\cdot)$ denote the Kronecker product and vectorization operator respectively (Magnus et al., 1995). Denote $\mathbf{D} = \left\{\text{vec}(\mathbf{D}^1), \cdots, \text{vec}(\mathbf{D}^r)\right\}_{q(p+1)\times r}$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_r)^{\mathrm{T}}$. Then (5) can be written as

$$\mathbf{Y}_i = (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}\boldsymbol{\beta}(T_i) + \mathbf{E}_i. \tag{7}$$

We need to estimate the r principal component functions stored in the vector $\boldsymbol{\beta}(T_i)$ and the $(p+1)qr$ unknown parameters in $\mathbf{D}$.

For estimation of the unknown principal component functions, we approximate them using polynomial splines and apply the penalized least squares estimation with a sparsity-inducing penalty function. Specifically, let $\mathbf{b}(t) = \{b_1(t), \cdots, b_K(t)\}^{\mathrm{T}}$ be a spline basis with dimension K. For the l-th principal component function $\beta_l(t)$, we write $\beta_l(t) \approx \mathbf{a}_l^{\mathrm{T}}\mathbf{b}(t)$, where $\mathbf{a}_l$ is the spline coefficient vector in the basis expansion. Then we have $\boldsymbol{\beta}(t) \approx \mathbf{A}^{\mathrm{T}}\mathbf{b}(t)$, where $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_r)$ is the $K \times r$ matrix of spline coefficients. Ignoring the spline approximation error in the coefficient matrix estimation, the reduced model (5) then takes the form of

$$\mathbf{Y}_i = (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}\mathbf{A}^{\mathrm{T}}\mathbf{b}(T_i) + \mathbf{E}_i,$$

with the identifiability constraints

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r, \ \int_{\mathcal{T}} \mathbf{b}(t)\mathbf{b}^{\mathrm{T}}(t) = \mathbf{I}_K, \tag{8}$$

The constraints in (8) imply that $\int_{\mathcal{T}} \boldsymbol{\beta}(t)\boldsymbol{\beta}^{\mathrm{T}}(t) = \mathbf{I}_r$, which is a usual orthogonality constraint imposed on the principal component functions.

When p is large, the reduction through principal component is not sufficient to obtain a parsimonious model. This can be seen from the fact that when the model is linear, the dimension reduction above becomes void and the estimation still suffers from high dimensionality. However, in applications, many covariates are irrelevant (or almost so) for prediction of the responses; see for example Ma et al. (2015). This fact opens the door for reaching a more parsimonious model through selection of relevant covariates. The selection can be achieved by introducing sparsity-inducing penalization to the method of least squares.

To add a proper penalty function, we take a further look at the $q(p+1) \times r$ coefficient matrix $\mathbf{D}$ in (7). We consider the row-wise partition $\mathbf{D} = (\mathbf{D}_0^{\mathrm{T}}, \cdots, \mathbf{D}_p^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{D}_j \in \mathbb{R}^{q\times r}$, $0 \le j \le p$, where the sub-matrix block $\mathbf{D}_j$ contains all unknown coefficients associated with the j-th predictor. Moreover, the (l, v)-th entry in $\mathbf{D}_j$ is $d_{jl,v}$, the principal component loading of β_v on the coefficient $f_j^{(l)}$ corresponding to the l-th response and j-th predictor. In light of (4), when $\mathbf{D}_j = \mathbf{0}$, the coefficient functions $f_j^{(l)}$ become a zero function for all responses. In other words, the j-th predictor is irrelevant for all responses simultaneously if and only if $\mathbf{D}_j = \mathbf{0}$. Thus we consider the penalized least square problem

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{Y}_i - (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}\mathbf{A}^{\mathrm{T}}\mathbf{b}(T_i)\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\mathbf{D}_j\|), \tag{9}$$

where $\|\cdot\|$ for a matrix denotes its Frobenius norm, $p_\lambda(\cdot)$ is a penalty function and λ is a penalty parameter. The resulting estimator for $f_j^{(l)}(t)$ becomes $\widehat{f}_j^{(l)}(t) = \sum_{k=1}^{K} \widehat{c}_{jk}^{(l)} b_k(t)$, where $\widehat{c}_{jk}^{(l)} = \sum_{v=1}^{r} \widehat{a}_{kv}\widehat{d}_{jk,v}$. Although other sparsity-inducing penalties can be used, in this paper we use the Smoothly Clipped Absolute Deviation (SCAD) penalty function (Fan and Li, 2001), defined by

(10)

where a is a parameter which is typically set to be 3.7 according to Fan and Li (2001). Our application of a penalty on a matrix norm is similar in spirit to the group-wise penalization (Yuan and Lin, 2006; Huang et al., 2012; Gong et al., 2013). Our situation differs to existing work in two aspects: first, our group is a matrix instead of a vector; second, we need to minimize an objective function with respect to an orthonormal matrix $\mathbf{A}$.

The solution to the optimization problem (9) is not unique. Suppose $(\widehat{\mathbf{D}}, \widehat{\mathbf{A}})$ is a solution to (9) and $\mathbf{Q}$ is an $r \times r$ orthogonal matrix. Let $\widetilde{\mathbf{D}} = \widehat{\mathbf{D}}\mathbf{Q}$ and $\widetilde{\mathbf{A}} = \widehat{\mathbf{A}}\mathbf{Q}$, then $\widetilde{\mathbf{D}}_j = \widehat{\mathbf{D}}_j\mathbf{Q}$. Since $\|\widehat{\mathbf{D}}_j\| = \|\widetilde{\mathbf{D}}_j\|$, $\widetilde{\mathbf{A}}^{\mathrm{T}}\widetilde{\mathbf{A}} = \mathbf{I}_r$, and $\widetilde{\mathbf{D}}\widetilde{\mathbf{A}}^{\mathrm{T}} = \widehat{\mathbf{D}}\widehat{\mathbf{A}}^{\mathrm{T}}$, $(\widetilde{\mathbf{D}}, \widetilde{\mathbf{A}})$ is also a solution to (9). Similar to Chen and Huang (2012), we have the following result.

Lemma 1 *The solution* $(\mathbf{D}, \mathbf{A})$ *to the optimization problem* (9) *is unique up to an* $r \times r$ *orthogonal matrix.*

This lemma says that, although $(\mathbf{D}, \mathbf{A})$ is not identified but $\mathbf{C} := \mathbf{D}\mathbf{A}^{\mathrm{T}}$ is. Moreover, the coefficient function matrix $\mathbf{F}$ is identifiable. To see this, note that $\mathrm{vec}(\mathbf{F}) = \mathbf{D}\boldsymbol{\beta}$, with the spline approximation that $\boldsymbol{\beta} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$, we have that $\mathrm{vec}(\mathbf{F}) = \mathbf{D}\mathbf{A}^{\mathrm{T}}\mathbf{b} = \mathbf{C}\mathbf{b}$ is identifiable.

Since $\mathrm{rank}(\mathbf{A}) = r$, $\mathrm{rank}(\mathbf{C}) \le r$. On the other hand, if $\mathrm{rank}(\mathbf{C}) \le r$, the QR decomposition $\mathbf{C} = \mathbf{D}\mathbf{A}^{\mathrm{T}}$ will give us $(\mathbf{D}, \mathbf{A})$. Assuming $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r$ and writing $\mathbf{C} = (\mathbf{C}_0^{\mathrm{T}}, \cdots, \mathbf{C}_p^{\mathrm{T}})^{\mathrm{T}}$, we have $\mathbf{C}_j = \mathbf{D}_j\mathbf{A}^{\mathrm{T}} \in \mathbb{R}^{q\times K}$, and $\|\mathbf{C}_j\| = \|\mathbf{D}_j\|$. Therefore, the optimization problem (9) is equivalent to a rank constrained optimization problem with respect to $\mathbf{C}$:

$$\min_{\mathbf{C}:\,\mathrm{rank}(\mathbf{C})\le r} \sum_{i=1}^{n} \|\mathbf{Y}_i - (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{C}\mathbf{b}(T_i)\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\mathbf{C}_j\|). \tag{11}$$

This equivalent representation of (9) is convenient when we study the asymptotic properties of our estimator.

2.2 Computational Algorithm

We solve (9) by using block-wise coordinate proximal/gradient descent, i.e., alternatively updating $\mathbf{D}$ and $\mathbf{A}$ in a cyclic manner with proximal descent and gradient descent. To facilitate our

discussion, denote the least squares part and the penalty part of the objective function as

$$H(\mathbf{D},\mathbf{A}) = \sum_{i=1}^{n} ||\mathbf{Y}_i - (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}\mathbf{A}^{\mathrm{T}}\mathbf{b}(T_i)||^2 \quad \text{and} \quad G(\mathbf{D}) = n\sum_{j=0}^{p} P_\lambda(||\mathbf{D}_j||),$$

respectively. When $\mathbf{A}$ is fixed at $\widehat{\mathbf{A}}$, we denote $H_{\widehat{\mathbf{A}}}(\mathbf{D}) = H(\mathbf{D},\widehat{\mathbf{A}})$. Similarly, when $\mathbf{D}$ is fixed at $\widehat{\mathbf{D}}$, denote $H_{\widehat{\mathbf{D}}}(\mathbf{A}) = H(\widehat{\mathbf{D}},\mathbf{A})$.

For updating the $\mathbf{D}$ with the $\mathbf{A}$ fixed at $\mathbf{A}^{(k)}$, where k represents the iteration number, the objective function becomes

$$H_{\mathbf{A}^{(k)}}(\mathbf{D}) + G(\mathbf{D}). \tag{12}$$

There is no closed-form to solve $\mathbf{D}$ in (12). We thus consider the following proximal regularization function (Beck and Teboulle, 2009) of (12) at $\mathbf{D}^{(k)}$

$$H_{\mathbf{A}^{(k)}}(\mathbf{D}^{(k)}) + \langle \nabla H_{\mathbf{A}^{(k)}}(\mathbf{D}), \mathbf{D} - \mathbf{D}^{(k)} \rangle + \frac{1}{2\tau_{\mathbf{D}}^{(k+1)}} ||\mathbf{D} - \mathbf{D}^{(k)}||^2 + G(\mathbf{D}), \tag{13}$$

where $\tau_{\mathbf{D}}^{(k+1)} > 0$ is the step-size and $\nabla H_{\mathbf{A}^{(k)}}(\mathbf{D})$ has a closed-form:

$$\nabla H_{\mathbf{A}^{(k)}}(\mathbf{D}) = -2\sum_{i=1}^{n} (\mathbf{X}_i \otimes \mathbf{I}_q)\left\{\mathbf{Y}_i - (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}(\mathbf{A}^{(k)})^{\mathrm{T}}\mathbf{b}(T_i)\right\}\mathbf{b}(T_i)^{\mathrm{T}}\mathbf{A}^{(k)}. \tag{14}$$

$\mathbf{D}^{(k+1)}$ is defined to be the minimizer of (13), or equivalently, by completing the squares,

$$\mathbf{D}^{(k+1)} = \arg\min_{\mathbf{D}} ||\mathbf{D} - \{\mathbf{D}^{(k)} - \tau_{\mathbf{D}}^{(k+1)} \nabla H_{\mathbf{A}^{(k)}}(\mathbf{D})\}||^2/2 + \tau_{\mathbf{D}}^{(k+1)} G(\mathbf{D}). \tag{15}$$

This is the proximal operator (Beck and Teboulle, 2009) on $G(\mathbf{D})$ with a general step-size $\tau_{\mathbf{D}}$

$$\mathrm{Prox}_{\tau_{\mathbf{D}} G}(\mathbf{\Xi}) = \arg\min_{\mathbf{D}} \left\{\tau_{\mathbf{D}}\, G(\mathbf{D}) + (1/2)||\mathbf{D} - \mathbf{\Xi}||^2\right\}. \tag{16}$$

Following the idea of vector regularization problems such as Huang et al. (2012), the matrix form of the proximal operator on $G(\mathbf{D}) = n\sum_j P_\lambda(||\mathbf{D}_j||)$ with $P_\lambda(\cdot)$ defined in (10) has an analytical expression. In particular, denote $\mathrm{Prox}_{\tau_{\mathbf{D}} G}(\mathbf{\Xi})_j$, $j = 0, \cdots, p$, as the j-th $q \times r$ sub-matrix of

$\text{Prox}_{\tau_{\mathbf{D}}G}(\boldsymbol{\Xi})$ and let $\mathbb{S}(\boldsymbol{\Xi}_j;\lambda_j)$ be the soft-thresholding rule for the $q\times r$ matrix $\boldsymbol{\Xi}_j$ at level λ_j, i.e.,

$$\mathbb{S}(\boldsymbol{\Xi}_j;\lambda_j)=\begin{cases}0, & \text{if } \|\boldsymbol{\Xi}_j\|<\lambda_j,\\ \boldsymbol{\Xi}_j-\lambda_j\frac{\boldsymbol{\Xi}_j}{\|\boldsymbol{\Xi}_j\|}, & \text{otherwise.}\end{cases}$$

Then

$$\text{Prox}_{\tau_{\mathbf{D}}G}(\boldsymbol{\Xi})_j=\begin{cases}\mathbb{S}(\boldsymbol{\Xi}_j;\tau_{\mathbf{D}}n\lambda), & \text{if } \|\boldsymbol{\Xi}_j\|\le 2\,\tau_{\mathbf{D}}n\lambda,\\ \frac{a-1}{a-2}\mathbb{S}(\boldsymbol{\Xi}_j;\frac{a\,\tau_{\mathbf{D}}n\lambda}{a-1}), & \text{if } 2\,\tau_{\mathbf{D}}n\lambda<\|\boldsymbol{\Xi}_j\|\le a\,\tau_{\mathbf{D}}n\lambda,\\ \boldsymbol{\Xi}_j, & \text{if } \|\boldsymbol{\Xi}_j\|>a\,\tau_{\mathbf{D}}n\lambda.\end{cases} \tag{17}$$

Thus, the updating rule for $\mathbf{D}$ under proximal descent, when $\mathbf{A}$ is at $\mathbf{A}^{(k)}$, is

$$\mathbf{D}^{(k+1)}=\text{Prox}_{\tau_{\mathbf{D}}^{(k+1)}G}\{\mathbf{D}^{(k)}-\tau_{\mathbf{D}}^{(k+1)}\nabla H_{\mathbf{A}^{(k)}}(\mathbf{D}^{(k+1)})\}, \tag{18}$$

with $\nabla H_{\mathbf{A}^{(k)}}(\cdot)$ defined in (14) and $\text{Prox}_{\tau_{\mathbf{D}}^{(k+1)}G}(\cdot)$ defined in (17) respectively.

To choose the step-size $\tau_{\mathbf{D}}^{(k+1)}$ in the updating rule (18), one commonly uses a backtracking method to find one value of $\tau_{\mathbf{D}}^{(k+1)}$ such that the objective function monotonically decreases with steps. Similar to Gong et al. (2013), we take $\tau_{\mathbf{D}}^{(k+1)}=(1/2)^{\rho}$, where ρ is the minimal value among $\rho=1,2,\cdots$, such that the following criterion holds:

$$\begin{aligned}&H_{\mathbf{A}^{(k)}}(\mathbf{D}^{(k+1)})+G(\mathbf{D}^{(k+1)})\\ &\quad\le H_{\mathbf{A}^{(k)}}(\mathbf{D}^{(k)})+G(\mathbf{D}^{(k)})-\epsilon_{\mathbf{D}}^{(k+1)}\frac{1}{\tau_{\mathbf{D}}^{(k+1)}}\|\mathbf{D}^{(k+1)}-\mathbf{D}^{(k)}\|^2,\end{aligned} \tag{19}$$

where $\epsilon_{\mathbf{D}}$ is a fixed small number in $(0,1)$. A condition to ensure the boundedness of the step-size will be given at subsection 2.4.

Next, fixing $\mathbf{D}$ at $\mathbf{D}^{(k+1)}$, the objective function becomes $H_{\mathbf{D}^{(k+1)}}(\mathbf{A})$, with $\mathbf{A}$ satisfying the orthonormal constraint. The set of all $K\times r$ orthonormal matrices, denoted as $\text{St}(r,K)$ ($r\le K$) is called (orthogonal) Stiefel manifold (Edelman et al., 1998), i.e.,

$$\text{St}(r,K):=\{\mathbf{A}\in\mathbb{R}^{K\times r}:\mathbf{A}^{\mathrm{T}}\mathbf{A}=\mathbf{I}_r\}.$$

We use that fact that $\mathrm{St}(r, K)$ is an embedded sub-manifold of the $K \times r$ Euclidean space $\mathbb{R}^{K\times r}$ and update $\mathbf{A}^{(k+1)}$ by the gradient descent method (Absil et al., 2009; Edelman et al., 1998). The gradient descent updating rule for a sub-manifold includes four steps: first, we calculate the negative gradient of the objective function without any constraint in the Euclidean space; second, we find the tangent space of the sub-manifold structure at the current point, then project the negative gradient function of the Euclidean space onto the tangent space; third, calculate the value of updating $\mathbf{A}$ along the direction of projected negative gradient with given step-size $\tau_{\mathbf{A}}$; finally, retract the calculated value in the third step back to the manifold structure.

Now we specialize the four steps to our problem. With the objective function $H_{\mathbf{D}^{(k+1)}}(\mathbf{A})$, the gradient function on the Euclidean space, with respect to $\mathbf{A}$ is

$$\nabla H_{\mathbf{D}^{(k+1)}}(\mathbf{A}) = -2\sum_{i=1}^{n} \mathbf{b}(T_i)\left\{\mathbf{Y}_i - (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}^{(k+1)}\mathbf{A}^{\mathrm{T}}\mathbf{b}(T_i)\right\}^{\mathrm{T}} (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{D}^{(k+1)}. \tag{20}$$

The tangent space of $\mathrm{St}(r, K)$ at $\mathbf{A}$ can be explicitly expressed as

$$T_{\mathbf{A}}\mathrm{St}(r, K) = \{\mathbf{Z} \in \mathbb{R}^{K\times r} : \mathbf{A}^{\mathrm{T}}\mathbf{Z} + \mathbf{Z}^{\mathrm{T}}\mathbf{A} = 0\}.$$

The gradient function, from the view of Stiefel manifold structure, is

$$\mathrm{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}) = \mathrm{P}_{\mathbf{A}}\left\{\nabla H_{\mathbf{D}^{(k+1)}}(\mathbf{A})\right\}, \tag{21}$$

and the projection $\mathrm{P}_{\mathbf{A}}$ onto the tangent space $T_{\mathbf{A}}\mathrm{St}(r, K)$ has a closed-form

$$\mathrm{P}_{\mathbf{A}}\boldsymbol{\xi} = (\mathbf{I} - \mathbf{A}\mathbf{A}^{\mathrm{T}})\boldsymbol{\xi} + \mathbf{A}\,\mathrm{skew}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\xi}), \tag{22}$$

where $\mathrm{skew}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\xi}) = 1/2(\mathbf{A}^{\mathrm{T}}\boldsymbol{\xi} - \boldsymbol{\xi}^{\mathrm{T}}\mathbf{A})$. Denote $\mathrm{qf}(\boldsymbol{\xi})$ as the Q factor of the QR decomposition of $\boldsymbol{\xi} \in \mathbb{R}^{K\times r}$, which retracts $\boldsymbol{\xi}$ back to the manifold $\mathrm{St}(r, K)$. When $\mathbf{D}$ is fixed at $\mathbf{D}^{(k+1)}$, the gradient descent method on the Stiefel manifold updates $\mathbf{A}$ by

$$\mathbf{A}^{(k+1)} = \mathrm{qf}[\mathbf{A}^{(k)} - \tau_{\mathbf{A}}^{(k+1)}\,\mathrm{P}_{\mathbf{A}^{(k)}}\{\nabla H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})\}], \tag{23}$$

where $\tau_{\mathbf{A}}^{(k+1)} > 0$ is the step-size, and $\mathrm{P}_{\mathbf{A}^{(k)}}$ and $\nabla H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})$ are defined above. The step-size $\tau_{\mathbf{A}}^{(k+1)}$ can be chosen by the Armijo backtracking method (Absil et al., 2009, Chapter 4.2). In particular,

$\tau_{\mathbf{A}}^{(k+1)} = (1/2)^{\rho}$ where ρ is the minimal value among $\rho = 1, 2, \cdots$, such that the following criterion holds:

$$H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k+1)}) \leq H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)}) - \epsilon_{\mathbf{A}}\tau_{\mathbf{A}}^{(k+1)}\|\text{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})\|^2, \tag{24}$$

with some fixed $\epsilon_{\mathbf{A}} \in (0, 1)$. In the updating formula (23), the projection $\mathrm{P}_{\mathbf{A}^{(k)}}$ is critical to ensure that the gradient is correct, and the retraction guarantees the minimizer is on the manifold.

The above discussion leads to Algorithm I. The convergence property of the algorithm will be discussed in subsection 2.4. In the inner loops of this algorithm, we only iterate for **D** and **A** once. In computer implementation, one can also iterate multiple times in the inner loop; this can help reduce the number of outer loops at the cost of more computational time for the inner loops.

Algorithm I: Algorithm Using Coordinate Proximal/Gradient Descent

Input : $\mathbf{Y} \in \mathbb{R}^{q\times n}$, $\mathbf{X} \in \mathbb{R}^{(p+1)\times n}$, $\mathbf{T} \in \mathbb{R}^{n}$, $\lambda > 0$, $r \geq 1$, $\epsilon > 0$.
Output: $\widehat{\mathbf{D}}(\widehat{\mathbf{A}})^{\mathrm{T}}$.
for *k from* $0, 1, \cdots$ **do**

- Fixed $\mathbf{A}^{(k)}$, update $\mathbf{D}^{(k+1)}$ by (18),
  with the step-size $\tau_{\mathbf{D}}$ chosen by backtracking method (19); ;
- Fixed $\mathbf{D}^{(k+1)}$, update $\mathbf{A}^{(k+1)}$ by (23), ;
  the step-size $\tau_{\mathbf{A}}$ is chosen by Amijo backtracking method (24); ;
- Check whether the following stopping criterion is satisfied ;
  $H(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) + G(\mathbf{D}^{(k)}) - H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k+1)}) - G(\mathbf{A}^{(k+1)}) < \varepsilon$.

end

2.3 Tuning Parameters

The tuning parameters, including the number of spline basis, the penalty parameter λ, and the number of components r, can be determined using the K-fold cross-validation (CV). We found that the typical 5-fold cross-validation does not produce stable results. We used 10-fold CV in the simulation studies and 50-fold CV in the real data application, which produced stable results.

2.4 Convergence Analysis of the Algorithm

Two features of the optimization problem (9) makes convergence analysis of the algorithm challenging: (i) the objective function is non-convex, and (ii) the manifold structure on $\mathbf{A}$ imposes non-convex constraints that the optimization algorithm needs to respect. The non-convexity makes it difficult to ensure an algorithm to converge to the global optimal. Moreover, when applying a numerical method for optimization on manifold structure, such as Stiefel manifold, even convergence to a local optimal cannot be guaranteed (Absil et al., 2009). In this subsection, we show that every accumulation point of the parameter sequence generated by Algorithm I is a critical point of (9). Here, the critical point refers to a point whose gradient is zero or the sub-gradient contains zero. Convergence to a critical point is a necessary but not sufficient condition for convergence to a local optimal. Although this kind of result is weaker than we desire, similar results are typically seen for complex optimization problems involved in statistics and applied mathematics, such as Yun et al. (2011) for coordinate gradient descent method, Bunea et al. (2012) for variable selection under low rank constraint, and Wang et al. (2015) for ADMM.

Now we give a precise definition of critical point in our context. Denote $\partial G/\partial \mathbf{D}$ as the sub-gradient of G with respect to $\mathbf{D}$. It has a closed-form such that restricted on the j-th $q\times r$ sub-matrix $\mathbf{D}_j$, $j = 0, \cdots, p$,

$$\frac{\partial G}{\partial \mathbf{D}}\Big|_j = \begin{cases} \mathbf{0}, & \text{if } \alpha\lambda \le \|\mathbf{D}_j\|, \\ \frac{-2\,\mathbf{D}_j + 2\,\alpha\lambda\frac{\mathbf{D}_j}{\|\mathbf{D}_j\|}}{2\,(\alpha-1)}, & \text{if } \lambda < \|\mathbf{D}_j\| \le \alpha\lambda, \\ \lambda\frac{\mathbf{D}_j}{\|\mathbf{D}_j\|}, & \text{if } 0 < \|\mathbf{D}_j\| \le \lambda, \\ \overline{B(\mathbf{0},1)}, & \text{if } \|\mathbf{D}_j\| = 0, \end{cases} \tag{25}$$

where $B(\mathbf{0}, 1)$ is the unit ball on $\mathbb{R}^{q\times r}$ centered at $\mathbf{0}$.

Definition 1 $(\mathbf{D}^*, \mathbf{A}^*) \in \mathbb{R}^{(p+1)q\times r} \times \mathrm{St}(r, K)$ *is said to be a critical point of* $H(\mathbf{D}, \mathbf{A}) + G(\mathbf{D})$ *if*

$$\mathbf{0}_{(p+1)q\times r} \in \frac{\partial H(\mathbf{D}^*, \mathbf{A}^*)}{\partial \mathbf{D}} + \frac{\partial G(\mathbf{D}^*)}{\partial \mathbf{D}} \tag{26}$$

and

$$\mathbf{0}_{K\times r} = \mathrm{P}_{\mathbf{A}^*}\left\{\frac{\partial H(\mathbf{D}^*, \mathbf{A}^*)}{\partial \mathbf{A}}\right\}, \tag{27}$$

where $\partial H(\mathbf{D}^*, \mathbf{A}^*)/\partial \mathbf{D} = \nabla H_{\mathbf{A}^*}(\mathbf{D}^*)$, $\partial H(\mathbf{D}^*, \mathbf{A}^*)/\partial \mathbf{A} = \nabla H_{\mathbf{D}^*}(\mathbf{A}^*)$, $\mathrm{P}_{\mathbf{A}^*}$ *and* $\partial G/\partial \mathbf{D}$ *are defined in* (14), (20), (22) *and* (25) *respectively.*

We need the following regularity condition.

(C0) $\partial H(\mathbf{D}, \mathbf{A})/\partial \mathbf{D}$ is uniformly Lipschitz continuous, i.e., there exits a positive constant L, such that

$$\left\| \frac{\partial H(\mathbf{D}, \mathbf{A})}{\partial \mathbf{D}} - \frac{\partial H(\mathbf{\Xi}, \mathbf{A})}{\partial \mathbf{\Xi}} \right\| \leq L \|\mathbf{D} - \mathbf{\Xi}\|$$

for $\forall\, \mathbf{\Xi}, \mathbf{D} \in \mathbb{R}^{(p+1)q \times r}$ and $\mathbf{A} \in \mathrm{St}(r, K)$.

This condition is satisfied if the largest eigenvalue of $\mathrm{Hess}(H)$ is uniformly bounded, where $\mathrm{Hess}(H)$ is the Hessian operator of function H. It is used to guarantee the boundedness of step-size $\tau_{\mathbf{D}}$.

Theorem 1 *Assume the condition* (C0) *holds, then all accumulation points of the sequence* $\{(\mathbf{D}^{(k)}, \mathbf{A}^{(k)})\}$ *generated by Algorithm* I *are critical points of* (9).

Note that the objective function (9) is nonnegative. Since we check convergence by looking at the value of the objective function and each outer loop iteration reduces the objective function by at least $\epsilon > 0$, Algorithm I stops in finite number of steps. The proof of Theorem 1 is given in Appendix A.

2.5 Asymptotic Analysis

In this subsection, we study the asymptotic behavior of the estimator when the sample size n goes to infinity. Denote s to be the number of relevant covariates. We allow p, q, s, K and r ($r \leq K$) to grow with n.

Without loss of generality, let $\mathcal{T} = [0, 1]$ and for notational convenience, we let the relevant covariates to be the intercept and the first s predictor variables. We use C to refer to a generic constant that may change values from context to context. We need the following regularity conditions.

(C1) The index variable T has a continuous density supported on $[0,1]$ and the density is bounded away from zero and infinity on $[0,1]$.

(C2) $|X_j| \le C$ *a.s.*, $0 \le j \le p$. Moreover, the eigenvalues of $\mathbb{E}[\mathbf{X}_{0:s}\mathbf{X}_{0:s}^{\mathrm{T}}|T=t]$ are bounded away from zero and infinity, uniformly for $t \in [0,1]$, where $\mathbf{X}_{0:s} = (X_0, \ldots, X_s)^{\mathrm{T}}$.

(C3) The noise matrix $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)^{\mathrm{T}}$ has independent and identically distributed rows, with the vector $\mathbf{E}_i$ being sub-Gaussian in the sense that its moment-generating function satisfies $\mathbb{E}\exp(t\mathbf{E}_i^{\mathrm{T}}\boldsymbol{\eta}) \le \exp(Ct^2\|\boldsymbol{\eta}\|^2)$ for any $\boldsymbol{\eta} \in \mathbb{R}^q$.

(C4) The rows of the true parameters $\mathbf{D}_0$ in model (7) has Euclidean norms bounded by a constant.

(C5) For $g = \beta_{0\nu}$, $1 \le \nu \le r$, where $\beta_{0\nu}$'s denote the true principal component functions, 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 biggest 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 $m \ge d + 1/2$.

These assumptions are commonly used in the literature on sparse nonparametric models. In particular, (C5) implies the L_2 norm of $f_{0j}^{(l)} = \sum_{\nu=1}^{r} d_{0jl,\nu}\beta_{0\nu}$ is bounded, where $d_{0jl,\nu}$ is the $(jq+l,\nu)$-th entry of the true parameter matrix $\mathbf{D}_0$. Under the smoothness assumption for $\beta_{0\nu}$ as given in Condition (C5), there exists $\mathbf{a}_{0\nu} = (a_{01\nu}, \ldots, a_{0K\nu})^{\mathrm{T}} \in \mathbb{R}^K$ such that

$$\widetilde{\beta}_{0\nu}(t) = \sum_{k=1}^{K} a_{0k\nu} b_k(t), \ \|\beta_{0\nu} - \widetilde{\beta}_{0\nu}\|_\infty = O(K^{-d}), \tag{28}$$

where $\|\beta\|_\infty = \sup_{t\in[0,1]}|\beta(t)|$ is the L_∞ norm of the function β.

Let $\mathbf{C}_0$ be the true parameter of $\mathbf{C}$, i.e., $\mathbf{C}_0 = (\mathbf{C}_{00}^{\mathrm{T}}, \ldots, \mathbf{C}_{0p}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{D}_0\mathbf{A}_0^{\mathrm{T}}$, $\mathbf{A}_0 = (\mathbf{a}_{01}, \cdots, \mathbf{a}_{0r})$, and let $a_n \ll b_n$ mean $a_n = o(b_n)$ for positive numbers b_n.

Theorem 2 *(Convergence rates for estimation of* $\mathbf{C}_0$*) Assume Conditions* (C1)–(C5) *hold,* $K \to \infty$, $Ks^2\log(Ks)/n \to 0$, *and*

$$\frac{Kq\log(Kpq)}{n} + \frac{r(K+sq-r)}{n} + \frac{sq}{K^{2d}} \ll \lambda^2 \ll \min_{j\le s}\sum_{l=1}^{q}\|f_{0j}^{(l)}\|^2. \tag{29}$$

Then there is a local minimizer $\widehat{\mathbf{C}}$ *of* (11) *that satisfies*

i. $\mathbb{P}(\widehat{\mathbf{C}}_j \equiv \mathbf{0}, j > s) \to 1,$ *as* $n \to \infty,$

ii. $$\|\widehat{\mathbf{C}} - \mathbf{C}_0\|^2 = O_p\left(\frac{(sq + K - r)r}{n} + \frac{sq}{K^{2d}}\right). \tag{30}$$

Moreover, the estimated coefficient functions have the following convergence rate

$$\sum_{j=0}^{p}\sum_{l=1}^{q}\|\widehat{f}_j^{(l)} - f_{0j}^{(l)}\|^2 = O_p\left(\frac{(sq + K - r)r}{n} + \frac{sq}{K^{2d}}\right), \tag{31}$$

where $f_{0j}^{(l)}$ *is the true coefficient function,* $0 \le j \le p$, $1 \le l \le q$.

Recall that for notational convenience the relevant covariates are the intercept and the first s predictor variables in the statement of the theorem. Property (i) says that with probability goes to one, the irrelevant covariates will not be included in the selected model. This property indicates the consistency in variable selection and is sometimes referred to as the support recovery property. In the assumption (29) on λ, the lower bound is used to avoid over-fitting and the upper bound to guarantee support recovery.

The proof of the theorem in fact shows that the estimator defined with the knowledge that which s covariates are relevant also has the rate of convergence given in (31) and the irrelevant covariates can be consistently identified by the penalized least squares method. Therefore, our asymptotic result shows that the penalized estimator defined without knowledge of relevant covariates has the same convergence rate as when we know which covariates are irrelevant. Thus we can say the estimator has the nonparametric oracle property as defined in Storlie et al. (2011). This theorem allows s, q and r to vary with n. If they do not vary with n, the rate of convergence is $O_P(K/n + K^{-2d})$, the standard rate of convergence for the spline regression; and when $K_n \sim n^{1/(2d+1)}$, we obtain the well-known optimal rate of convergence $n^{-2d/(2d+1)}$ of Stone (1982).

Remark 1 *Theorem* 1 *showed that all accumulation points of the sequence generated by Algorithm* I *are critical points. Theorem* 2 *gives the asymptotic property of a local minimizer of* (11) *and corresponding function estimators. Since Theorem* 1 *does not imply that Algorithm* I *converges to a local minimum of the optimization problem, Theorem* 2 *can not be combined with Theorem* 1 *to show the statistical property of the estimator generated by Algorithm* I. *How to fill in this theoretical gap is an open question.*

2.6 Simulation Study

We conducted a simulation study to evaluate the proposed method. We considered two cases: i. the data are generated from a reduced MVCM model (5) where the coefficient functions are exactly spanned by principal component functions; ii. the data are generated from a MVCM where the coefficient functions are represented by principal component functions plus some random noises. In the second case, the MVCM is not in the reduced form (5). In the first case, we found that the K-fold CV can accurately select the true number of components and it can also identify the number of irrelevant covariates with high accuracy. In the second case, there does not exist a true fixed number of principal components, and the K-fold CV helped find a reduced MVCM that is a good approximation to the true model. In both cases, the reduced MVCM has clear advantages over the full MVCM and the linear model in estimation of functional coefficients and identification of relevant covariates. Details of the simulation setups and results are given in the "Supplementary Materials".

3 Framingham Heart Study

The Framingham Heart Study (FHS) (Dawber et al., 1951) is a project in health research to identify the common factors that contribute to cardiovascular diseases. We used a subset of the data with 325 patients, which have measurements on 15 phenotypes in addition to the SNP information. The 15 phenotypes are shown in Table 1, and were used as the response variables in our application of the MVCM. To remove the effects of big differences in the scales of the response variables, we normalized the response variables to have mean 0 and standard deviation 1. The index variable used in MVCM is the level of sedentary activity in term of hours per day. After matching the SNP data with the phenotypes and deleting observations with missing values, there are 292 patients remaining in our study. With a large number of SNPs (32164 SNPs for chromosome 6 that we focused on), clearly it is crucial to identify a small number of them that are important in explaining the response variables. Each SNP has three possible allele combinations coded as -1, 0,

1. For details on genotyping, see http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v20.p8.

To reduce the large number of SNPs to a manageable size, we first applied our method by treating one SNP as a predictor at a time to select the top 500 SNPs mostly related to the responses, where the R^2 is used to rank the SNPs. In this screening process, we fixed the spline order at $m = 4$ and used $K = 8$ B-splines as the basis. The selected 500 SNPs were used in the comparative study to be reported below.

We considered three methods: the reduced MVCM (rMVCM), the full MVCM (fMVCM) and the linear model (LINEAR). All three methods involved variable selection using the SCAD penalty. We used the 50-fold CV to select the penalty parameters. We also used the 50-fold CV to select K, the number of spline basis functions. The reason of using a large fold number of CV is to improve the stability of the penalty parameter selection. The least CV errors by the rMVCM, fMVCM and LINEAR are respectively $0.054, 0.177, 0.129$, while the corresponding R^2's are $0.11, 0.97, 0.80$. The CV error by rMVCM is more than 50% reduction of the CV errors of the other two. The much larger R^2's for fMVCM and LINEAR associated with larger CV errors indicate strong over-fitting of these two methods. As to variable selection, rMVCM selected 9 significant SNPs while fMVCM and LINEAR selected 119 and 417 respectively. This result is consistent with the observations in the simulation study that fMVCM and LINEAR tend to over-select significant SNPs.

To obtain a biological explanation on the selected SNPs, we input the submitted *ss*# of the selected SNPs to the NCBI database (Sherry et al., 2001) to retrieve the *rs*# records. Among the 9 selected SNPs by rMVCM, three SNPs have been scientifically confirmed. In particular, the SNP *rs*4896044 has association with hypertension (Burton et al., 2007), while the SNPs *rs*9479367 and *rs*9321440 have associations with multiple diseases or symptoms (Gagliardi, 2011; McElroy, 2013; Liu et al., 2013).

Figure 2 shows the estimated coefficient functions for the confirmed SNP *rs*4896044. We observe that the estimated coefficients show different nonlinear patterns changing with sedentary activity hours. For example, for ventricular.rate and qrs.angle, the estimated coefficients show a

decreasing pattern, and have a sharp drop when people have more than 15 sedentary activity hours. For systolic.blood.pressure and phy.sys.1st(2nd).bp.read, the effect of this SNP shows similar increasing patterns, changing from negative to positive values as the value of sedentary activity hours increases; this similarity in patterns could be explained by the high correlations among these three measurements. For diastolic blood pressure measurements, the effect is relative flat when the number of sedentary activity hours is more than 6. On the other hand, the effect of SNP $rs4896044$ on other variables such as weight, bi.deltoid.girth and waist.girth are relatively weak since the estimated coefficients are close to zero for all values of the sedentary hours.

4 Conclusion

This paper extends the widely-used varying-coefficient model to multivariate responses and with a large number of covariates. We developed a novel estimation method by employing the non-centered principal components to significantly reduce the number of unknown functions to be estimated and imposing sparsity-inducing penalization to automatically select the relevant covariates. The proposed method has a wide range of applications, and it is particularly useful to identify variables when their effects may be altered by another variable in high-dimensional settings, such as gene-environment interactions in genome-wide association studies (GWAS). Our method requires the response variables to be continuous. In real data applications, however, we may have discrete response variables, such as disease status. Thus how to incorporate both continuous and discrete response variables in the dimension reduction and variable selection procedure is an interesting future research topic. Moreover, FHS is a continuing project containing longitudinal observations, so extending the proposed method to longitudinal data settings is also of interest, and needs further investigation.

Acknowledgements

The authors thank the Editors, the Associate Editor, and two reviewers for their comments that helped significantly improve this work.

Appendix A: Proof of Theorem 1

To prove Theorem 1, we first show that the backtracking criterion (19) can be satisfied when $\tau_{\mathbf{D}}^{(k+1)}$ is small enough.

Lemma A.1 *Suppose* $\nabla H_{\mathbf{A}^{(k)}}(\mathbf{D})$ *is Lipschitz continuous, i.e., there exits a positive constant* L*, such that* $\|\nabla H_{\mathbf{A}^{(k)}}(\mathbf{\Xi}) - \nabla H_{\mathbf{A}^{(k)}}(\mathbf{D})\| \leq L\,\|\mathbf{\Xi} - \mathbf{D}\|$ *for* $\forall\ \mathbf{\Xi}, \mathbf{D} \in \mathbb{R}^{(p+1)q\times r}$*. Then the backtracking line search criterion* (19) *is satisfied whenever* $1/\tau_{\mathbf{D}}^{(k+1)} \geq L/(1-\epsilon_{\mathbf{D}})$.

The proof of this Lemma is similar as Lemma 1 and Lemma 2 of Gong et al. (2013) and is given in "Supplementary Materials".

We now give the oroof of Theorem 1. Based on the backtracking method in Algorithm I, $\{\mathbf{D}^{(k)}, \mathbf{A}^{(k)}\}$ satisfies $H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k+1)}) + G(\mathbf{D}^{(k+1)}) < H(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) + G(\mathbf{D}^{(k)})$ which implies the sequence $\{H(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) + G(\mathbf{D}^{(k)})\}_{k=0,1,\cdots}$ is monotonically decreasing. Let $(\mathbf{D}^*, \mathbf{A}^*)$ be a accumulation point of $\{\mathbf{D}^{(k)}, \mathbf{A}^{(k)}\}$, i.e., there is a subsequence κ such that

$$\lim_{k\in\kappa\to\infty} (\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) = (\mathbf{D}^*, \mathbf{A}^*). \tag{A.1}$$

Since $\mathrm{St}(r, K)$ is a compact subset of $\mathbb{R}^{K\times r}$, we have $\mathbf{A}^* \in \mathrm{St}(r, K)$, i.e., $(\mathbf{D}^*, \mathbf{A}^*) \in \mathbb{R}^{(p+1)q\times r} \times \mathrm{St}(r, K)$. Note that $H(\mathbf{D}, \mathbf{A}) + G(\mathbf{D})$ is bounded from below. Hence the monotonicity of $\{H(\mathbf{D}, \mathbf{A}) + G(\mathbf{D})\}_{k=0,1,\cdots}$ implies

$$\lim_{k\to\infty} H(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) + G(\mathbf{D}^{(k)}) = \lim_{k\in\kappa\to\infty} H(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) + G(\mathbf{D}^{(k)}) = H(\mathbf{D}^*, \mathbf{A}^*) + G(\mathbf{D}^*).$$

The backtracking rule in (19) and (24) imply that

$$\begin{aligned} & H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k+1)}) + G(\mathbf{D}^{(k+1)}) \\ \leq\ & H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k)}) - \epsilon_{\mathbf{A}}\tau_{\mathbf{A}}^{(k+1)}\|\mathrm{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})\|^2 + \Big\{H(\mathbf{D}^{(k)}, \mathbf{A}^{(k)}) \\ & - \epsilon_{\mathbf{D}}\,\frac{1}{\tau_{\mathbf{D}}^{(k+1)}}\|\mathbf{D}^{(k+1)} - \mathbf{D}^{(k)}\|^2 + G(\mathbf{D}^{(k)}) - H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k)})\Big\}. \end{aligned} \tag{A.2}$$

Under our assumption of Theorem 1, the condition of Lemma **A.1** is be satisfied, we thus have $\liminf_{k\to\infty} 1/\tau_{\mathbf{D}}^{(k+1)} > 0$. Using the proof of Theorem 4.3.1 in Absil et al. (2009), we also have

$\liminf_{k\to\infty} \tau_{\mathbf{A}}^{(k)} > 0$. Hence, by taking lim sup on both sides of (A.2) with $k \in \kappa$, we have

$$\begin{aligned}\left(\liminf_{k\in\kappa\to\infty} \frac{1}{\tau_{\mathbf{D}}^{(k+1)}}\right) \limsup_{k\in\kappa\to\infty} \|\mathbf{D}^{(k+1)} - \mathbf{D}^{(k)}\| + \left(\liminf_{k\in\kappa\to\infty} \tau_{\mathbf{A}}^{(k)}\right) \limsup_{k\in\kappa\to\infty} \|\mathrm{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})\| \\ \leq \limsup_{k\in\kappa\to\infty} \left\{ \frac{1}{\tau_{\mathbf{D}}^{(k+1)}} \|\mathbf{D}^{(k+1)} - \mathbf{D}^{(k)}\| + \tau_{\mathbf{A}}^{(k)} \|\mathrm{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})\| \right\} \leq 0.\end{aligned}$$

Thus, we have

$$\lim_{k\in\kappa\to\infty} \|\mathbf{D}^{(k+1)} - \mathbf{D}^{(k)}\| = 0 \quad \text{and} \quad \lim_{k\in\kappa\to\infty} \|\mathrm{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)})\| = 0. \tag{A.3}$$

It can be shown by (A.1) and (A.3) that

$$\lim_{k\in\kappa\to\infty} \mathbf{D}^{(k+1)} = \mathbf{D}^*. \tag{A.4}$$

The smoothness of the projection operator $\mathrm{P}_{\mathbf{A}}$ and H, together with (A.1), (A.3) and (A.4), imply that

$$\mathrm{P}_{\mathbf{A}^*}\left\{ \frac{\partial H(\mathbf{D}^*, \mathbf{A}^*)}{\partial \mathbf{A}} \right\} = \lim_{k\in\kappa\to\infty} \mathrm{P}_{\mathbf{A}^{(k)}}\left\{ \frac{\partial H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k)})}{\partial \mathbf{A}} \right\} = \lim_{k\in\kappa\to\infty} \mathrm{grad}\, H_{\mathbf{D}^{(k+1)}}(\mathbf{A}^{(k)}) = \mathbf{0},$$

which show (27). Finally, considering that the minimizer of $\mathbf{D}^{(k+1)}$ is also a critical point of the proximal regularization function (13), then $\mathbf{0}$ belongs to the sub-gradient of (13). We thus have

$$\mathbf{0} \in \frac{\partial H(\mathbf{D}^{(k+1)}, \mathbf{A}^{(k)})}{\partial \mathbf{D}} + \frac{1}{\tau_{\mathbf{D}}^{(k+1)}} (\mathbf{D}^{(k+1)} - \mathbf{D}^{(k)}) + \frac{G(\mathbf{D}^{(k+1)})}{\partial \mathbf{D}}.$$

Taking limit on both sides of above equation with $k \in \kappa$ and using (A.1), (A.4), continuity of ∂H and ∂G and Lemma **A.1**, we obtain (26) and complete the proof.

Appendix B: Proof of Theorems 2

In order to clearly present the proof, we define $\mathbf{Y} = (\mathbf{Y}_1^{\mathrm{T}}, \ldots, \mathbf{Y}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{c} = \mathrm{vec}(\mathbf{C})$, $\mathbf{Z}^i = (\mathbf{b}(T_i) \otimes \mathbf{X}_i \otimes \mathbf{I}_q)$ and $\mathbf{Z} = (\mathbf{Z}^1, \ldots, \mathbf{Z}^n)^{\mathrm{T}}$. Then the least square part of (11) can be written as:

$$\sum_i^n \|\mathbf{Y}_i - (\mathbf{X}_i^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{C}\mathbf{b}(T_i)\|^2 = \|\mathbf{Y} - \mathbf{Z}\mathbf{c}\|^2. \tag{B.1}$$

We introduce some notations and additional definitions. In our proofs, C denotes a generic positive constant that might assume different values at different places. Recall the definition of the spline

coefficient matrix $\mathbf{C}_0$ given in Section 2.5, $\mathbf{C}_0$ can be written as $\mathbf{D}_0\mathbf{A}_0^{\mathrm{T}}$ with $\mathbf{D}_0 = (\mathbf{D}_{00}^{\mathrm{T}}, \ldots, \mathbf{D}_{0p}^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{D}_{0j}$ being $q \times r$ matrices and $\mathbf{A}_0$ a $K \times r$ matrix. Let $\mathbf{Z}_{0:s}$ be the $nq \times (s+1)qK$ sub-matrix of $\mathbf{Z}$ containing the columns corresponding to relevant covariates. More generally, we will use subscript $0:s$ to denote other sub-vectors/sub-matrices associated with irrelevant covariates.

Conditions (C1) and (C2) together imply that the eigenvalues of $\mathbf{Z}_{0:s}^{\mathrm{T}}\mathbf{Z}_{0:s}/n$ are also bounded away from zero and infinity, where $\mathbf{Z}_{0:s}$ contains the columns of $\mathbf{Z}$ associated with the first $s+1$ covariates, with probability approaching one.

We first present a result on the eigenvalues of $\mathbf{Z}_{0:s}^{\mathrm{T}}\mathbf{Z}_{0:s}$ which is used in various parts of the proof.

Lemma B.1 *Under the assumptions of Theorem 2, the eigenvalues of* $\mathbf{Z}_{0:s}^{\mathrm{T}}\mathbf{Z}_{0:s}/n$ *are bounded away from zero and infinity, with probability approaching one.*

The proof of this lemma is similar as Lemma A.2 of Huang et al. (2004) and is given is the "Supplementary Materials".

Now we are ready to give the proof of Theorem 2. The proof consists of two steps. Roughly speaking, we first show that the "oracle estimator" which assumes knowledge of zero blocks $\mathbf{C}_j$, $j > s$ achieves the convergence rate stated in Theorem 2 and then we show that this oracle estimator is actually a local minimizer of (11), which will complete the proof.

Formally, we define the oracle estimator as

$$\widetilde{\mathbf{C}}_{0:s} = \underset{\mathbf{C}_{0:s}:\mathrm{rank}(\mathbf{C}_{0:s})\le r}{\arg\min} \sum_i \|\mathbf{Y}_i - (\mathbf{X}_{i,0:s}^{\mathrm{T}} \otimes \mathbf{I}_q)\mathbf{C}_{0:s}\mathbf{b}(T_i)\|^2. \tag{B.2}$$

In the first part of the proof, we only consider the oracle estimator and we omit the subscript $0:s$ in the following for simplicity. As before, we can rewrite (B.2) as

$$\widetilde{\mathbf{c}} = \underset{\mathbf{c}=\mathrm{vec}(\mathbf{C}),\mathrm{rank}(\mathbf{C})\le r}{\arg\min} \|\mathbf{Y} - \mathbf{Z}\mathbf{c}\|^2. \tag{B.3}$$

Let $\mathbf{c}_0 = \mathrm{vec}(\mathbf{C}_0)$. Using that $\|\mathbf{Y} - \mathbf{Z}\widetilde{\mathbf{c}}\|^2 \le \|\mathbf{Y} - \mathbf{Z}\mathbf{c}_0\|^2$, and working out the squares, we obtain

$$\|\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0)\|^2 \le 2\langle \mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0), \mathbf{e} + \mathbf{R}\rangle,$$

where $\mathbf{e} = \mathrm{vec}(\mathbf{E})$ and $\mathbf{R} = \{\mathbf{X}_1^{\mathrm{T}}\mathbf{F}(T_1)^{\mathrm{T}}, \ldots, \mathbf{X}_n^{\mathrm{T}}\mathbf{F}(T_n)^{\mathrm{T}}\}^{\mathrm{T}} - \mathbf{Z}\mathbf{c}_0$ are the spline approximation errors.

Let $\Gamma = \{\boldsymbol{\eta} = \mathbf{Z}\mathbf{c}/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})} : \mathbf{c} = (\mathbf{c}_1^{\mathrm{T}}, \ldots, \mathbf{c}_K^{\mathrm{T}})^{\mathrm{T}}, \|\mathbf{c}\| \le 1, \mathrm{rank}\{(\mathbf{c}_1, \ldots, \mathbf{c}_K)\} \le r\}$, where $\lambda_{\max}(.)$ denotes the largest eigenvalue of a symmetric matrix. We first show that the covering entropy $\log N(\epsilon, \Gamma, l_2) \le r\{K + (s+1)q - r\}\log(C/\epsilon)$. In fact, for $\boldsymbol{\eta} = \mathbf{Z}\mathbf{c}/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})} \in \Gamma$, we can write $\mathbf{C} := (\mathbf{c}_1, \ldots, \mathbf{c}_K) = \mathbf{D}\mathbf{A}^{\mathrm{T}}$, $\mathbf{D} \in \mathbb{R}^{(s+1)q\times r}$, $\mathbf{A} \in \mathbb{R}^{K\times r}$, $\|\mathbf{D}\| \le 1$, $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$. The

covering number, under the Frobenius norm for $\mathbf{D}$ satisfying these assumptions is $(C/\epsilon)^{(s+1)qr}$. For the covering number of $\mathbf{A}$ satisfying the orthogonal condition, we use the distance $d(\mathbf{A}_1, \mathbf{A}_2) = ||\mathbf{A}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}||_{op}$, where $||.||_{op}$ denotes the operator norm, and then by Proposition 8 of Szarek (1982) the covering number is bounded by $(C/\epsilon)^{r(K-r)}$. Using that

$$\begin{aligned}||\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{D}_2\mathbf{A}_2^{\mathrm{T}}|| &\leq ||\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{D}_1\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}|| + ||\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_2\mathbf{A}_2^{\mathrm{T}} - \mathbf{D}_2\mathbf{A}_2^{\mathrm{T}}|| \\ &\leq ||\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}}||||\mathbf{A}_1\mathbf{A}_1^{\mathrm{T}} - \mathbf{A}_2\mathbf{A}_2^{\mathrm{T}}||_{op} + ||\mathbf{D}_1\mathbf{A}_1^{\mathrm{T}}\mathbf{A}_2 - \mathbf{D}_2||,\end{aligned}$$

we have $\log N(\epsilon, \Gamma, l_2) \leq r\{K + (s+1)q - r\}\log(C/\epsilon)$.

Furthermore, we have $\mathbb{E}\exp(t\langle\mathbf{e}, \boldsymbol{\eta}\rangle) \leq \exp(Ct^2||\boldsymbol{\eta}||^2)$ under Condition (C3). Using Dudley's integral entropy bound (for example see Theorem 3.1 of Koltchinskii (2011)), we get

$$\mathbb{E}\sup_{\boldsymbol{\eta}\in\Gamma}\langle\boldsymbol{\eta}, \mathbf{e}\rangle \leq C\int_0^2 \sqrt{r\{K + (s+1)q - r\}\log(\frac{C}{\epsilon})}d\epsilon \leq C\sqrt{r\{K + (s+1)q - r\}}.$$

The above implies that

$$\langle\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0)/\sqrt{\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})}, \mathbf{e}\rangle = ||\widetilde{\mathbf{c}} - \mathbf{c}_0|| \cdot O_p(\sqrt{r\{K + (s+1)q - r\}}),$$

which in turn implies

$$\langle\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0), \mathbf{e}\rangle = ||\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0)|| \cdot O_p(\sqrt{r\{K + (s+1)q - r\}}),$$

using Lemma **B.1**. Trivially, using Condition (C5), we have

$$\langle\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}), \mathbf{R}\rangle \leq ||\mathbf{R}||||\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c})|| = ||\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c})||O_p(\sqrt{n(s+1)qK^{-2d}}).$$

Thus

$$||\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0)||^2 = ||\mathbf{Z}(\widetilde{\mathbf{c}} - \mathbf{c}_0)|| \cdot O_p(\sqrt{r\{K + (s+1)q - r\}} + \sqrt{n(s+1)qK^{-2d}}) \tag{B.4}$$

which proved the convergence rate as in Theorem 2 (for the oracle estimator).

Now, we come to the second part of the proof where we show that the oracle estimator defined above is a local minimizer of the original problem (11) (without the information regarding zero rows of $\mathbf{C}$). Recall that we define the oracle estimator as $\widehat{\mathbf{c}}^o = \{(\widehat{\mathbf{c}}^o_{(1)})^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\}^{\mathrm{T}}$, where now we use $\widehat{\mathbf{c}}^o_{(1)}$

to denote the oracle estimator obtained by (B.3) using only the first s components of $\mathbf{X}_i$.

We first note that under our assumption $\lambda^2 \ll \min_{j\le s}\sum_{l=1}^{q}\|f_{0j}^{(l)}\|^2$ and the convergence rate presented in (B.4), $\|\widehat{\mathbf{c}}_j^o\| > a\lambda$ for $j \le s$, where $\widehat{\mathbf{c}}_j^o$ is the sub-vector of $\widehat{\mathbf{c}}^o$ associated with predictor j. Since $\|\widehat{\mathbf{c}}_j^o\| > a\lambda$ for $j \le s$ there is a small enough neighborhood of $\widehat{\mathbf{c}}_{(1)}^o$ such that $\|\mathbf{c}_{(1)j}\| > a\lambda$ for any $\mathbf{c}_{(1)}$ in this neighborhood, with rank of $\text{vec}^{-1}(\mathbf{c}_{(1)})$ bounded by r, where we use vec^{-1} to denote the operation of rearranging of a $(s+1)qK$-vector into a $(s+1)q \times K$ matrix, which implies $\sum_{j=0}^{s} p_\lambda(\|\widehat{\mathbf{c}}_j^o\|) = \sum_{j=0}^{s} p_\lambda(\|\mathbf{c}_{(1)j}\|)$. Thus, using that $\widehat{\mathbf{c}}^o$ is the oracle estimator, we have

$$\|\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{c}}^o\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\widehat{\mathbf{c}}_j^o\|) \le \|\mathbf{Y} - \mathbf{Z}\mathbf{c}\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\mathbf{c}_j\|), \tag{B.5}$$

for any $\mathbf{c}$ of the form $\mathbf{c} = (\mathbf{c}_{(1)}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{c}_{(1)}$ in a small neighborhood of $\widehat{\mathbf{c}}_{(1)}^o$ and $\text{rank}\{\text{vec}^{-1}(\mathbf{c}_{(1)})\} \le r$.

We will show that when $\delta > 0$ is sufficiently small, for any $\mathbf{c} = (\mathbf{c}_{(1)}^{\mathrm{T}}, \mathbf{c}_{(2)}^{\mathrm{T}})^{\mathrm{T}}$ that satisfies $\text{rank}\{\text{vec}^{-1}(\mathbf{c})\} \le r$, $\|\mathbf{c}_{(1)} - \widehat{\mathbf{c}}_{(1)}^o\| \le \delta$, $\|\mathbf{c}_{(2)}\| \le \delta$, the following inequality holds

$$\|\mathbf{Y} - \mathbf{Z}\widetilde{\mathbf{c}}\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\widetilde{\mathbf{c}}_j\|) \le \|\mathbf{Y} - \mathbf{Z}\mathbf{c}\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\mathbf{c}_j\|), \tag{B.6}$$

where $\widetilde{\mathbf{c}} = (\mathbf{c}_{(1)}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. Combining this inequality and (B.5) we obtain that the oracle estimator is a local minimum of the objective function (B.1).

It remains to show (B.6). The right hand side of (B.6) subtracting the left hand side is equal to $\|\mathbf{Z}_{(2)}\mathbf{c}_{(2)}\|^2 - 2\langle \mathbf{Y} - \mathbf{Z}_{(1)}\mathbf{c}_{(1)}, \mathbf{Z}_{(2)}\mathbf{c}_{(2)}\rangle + n\sum_{j=s+1}^{p} p_\lambda(\|\mathbf{c}_j\|)$, where $\mathbf{Z}_{(1)}$ contains the columns of $\mathbf{Z}$ associated with relevant covariates while $\mathbf{Z}_{(2)}$ contains the columns of $\mathbf{Z}$ associated with irrelevant covariates. We have $\|\mathbf{Z}_{(2)}\mathbf{c}_{(2)}\|^2 \le \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\mathbf{c}_{(2)}\|^2$. Furthermore,

$$\begin{aligned}|\langle \mathbf{Y} - \mathbf{Z}_{(1)}\mathbf{c}_{(1)}, \mathbf{Z}_{(2)}\mathbf{c}_{(2)}\rangle| &\le |\langle \mathbf{Z}_{(2)}^{\mathrm{T}}(\mathbf{Y} - \mathbf{Z}_{(1)}\widehat{\mathbf{c}}_{(1)}^o), \mathbf{c}_{(2)}\rangle| + |\langle \mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(1)}(\widehat{\mathbf{c}}_{(1)}^o - \mathbf{c}_{(1)}), \mathbf{c}_{(2)}\rangle| \\ &\le \widehat{\xi}\sum_{j=s+1}^{p}\|\mathbf{c}_j\| + \lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\mathbf{c}}_{(1)}^o - \mathbf{c}_{(1)}\|\|\mathbf{c}_{(2)}\|,\end{aligned}$$

where $\widehat{\xi} = \max_j \|\mathbf{Z}_j^{\mathrm{T}}(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{c}}^o)\|$, $\mathbf{Z}_j$ contains the columns of $\mathbf{Z}$ associated with predictor j, and we use $\lambda_{\max}(.)$ also to denote the largest singular value of a (non-symmetric) matrix. If δ is small enough, $p_\lambda(\|\mathbf{c}_j\|) = \lambda\|\mathbf{c}_j\|$ by the definition of the SCAD penalty. Thus, when $n\lambda - 2\widehat{\xi} \ge 0$,

$$\|\mathbf{Y} - \mathbf{Z}\mathbf{c}\|^2 + n\sum_{j=0}^{p} p_\lambda(\|\mathbf{c}_j\|) - \|\mathbf{Y} - \mathbf{Z}\widetilde{\mathbf{c}}\|^2 - n\sum_{j=0}^{p} p_\lambda(\|\widetilde{\mathbf{c}}_j\|)$$

$$\geq (n\lambda - 2\widehat{\xi}) \sum_{j>s} \|\mathbf{c}_j\| - 2\lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\mathbf{c}}_{(1)}^{o} - \mathbf{c}_{(1)}\|\|\mathbf{c}_{(2)}\| - \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\mathbf{c}_{(2)}\|^2$$

$$\geq (n\lambda - 2\widehat{\xi})\|\mathbf{c}_{(2)}\| - 2\lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\mathbf{c}}_{(1)}^{o} - \mathbf{c}_{(1)}\|\|\mathbf{c}_{(2)}\| - \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\mathbf{c}_{(2)}\|^2$$

$$\geq \|\mathbf{c}_{(2)}\|\{(n\lambda - 2\widehat{\xi}) - 2\lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\widehat{\mathbf{c}}_{(1)}^{o} - \mathbf{c}_{(1)}\| - \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\|\mathbf{c}_{(2)}\|\}$$

$$\geq \|\mathbf{c}_{(2)}\|\{(n\lambda - 2\widehat{\xi}) - 2\lambda_{\max}(\mathbf{Z}_{(1)}^{\mathrm{T}}\mathbf{Z}_{(2)})\delta - \lambda_{\max}(\mathbf{Z}_{(2)}^{\mathrm{T}}\mathbf{Z}_{(2)})\delta\}, \tag{B.7}$$

where we use the trivial inequality $\sum_{j>s}\|\mathbf{c}_j\| \geq \|\mathbf{c}_{(2)}\|$ in the third line above.

To get the order of $\widehat{\xi}$, note that $\widehat{\xi} = \max_j \|\mathbf{Z}_j^{\mathrm{T}}(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{c}}^{o})\| \leq \max_j\{\|\mathbf{Z}_j^{\mathrm{T}}\mathbf{e}\| + \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}(\widehat{\mathbf{c}}^{o} - \mathbf{c}_0)\|\}$. For the first part of the upper bound, note that $\mathbf{Z}_j$ is a $nq \times Kq$ matrix and the $(iq + l_1, kq + l_2)$-th entry of $\mathbf{Z}_j$ is $\delta_{l_1 l_2}\mathbf{X}_{ij}b_k(T_i)$, where $\delta_{l_1 l_2}$ is the Kronecker δ, $1 \leq i \leq n$, $1 \leq k \leq K$, $1 \leq l_1, l_2 \leq q$. To simplify the notation, we denote $\mathbf{Z}_{j,il_1}^{kl_2}$ as the $(iq + l_1, kq + l_2)$-th entry of $\mathbf{Z}_j$, and $\mathbf{Z}_j^{kl_2}$ as the $(kq + l_2)$-th row of $\mathbf{Z}_j$ respectively. We can write

$$\max_j \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{e}\| \leq \sqrt{Kq} \max_{0\leq j\leq p, 1\leq k\leq K, 1\leq l_2\leq q} \|(\mathbf{Z}_j^{kl_2})^{\mathrm{T}}\mathbf{e}\|.$$

The Kronecker product in the entry element $\mathbf{Z}_{j,il_1}^{kl_2}$ implies that only n out of nq entries in $\mathbf{Z}_j^{kl_2}$ are non-zero. Thus $(\mathbf{Z}_j^{kl_2})^{\mathrm{T}}\mathbf{e}$ is a sum of n non-zeros. This fact, together with Conditions (C2) and (C3), implies $\|(\mathbf{Z}_j^{kl_2})^{\mathrm{T}}\mathbf{e}\| = O_p(\sqrt{n})$. Applying Lemma 2.2.2 of van der Vaart and Wellner (1996), we have $\max_j \|\mathbf{Z}_j^{\mathrm{T}}\mathbf{e}\| = O_p(\sqrt{nKq\log(Kpq)})$. For the second part of upper bound, Condition (C3) implies $\|\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}(\widehat{\mathbf{c}}^{o} - \mathbf{c}_0)\| = \sqrt{n}\|\mathbf{Z}(\widehat{\mathbf{c}}^{o} - \mathbf{c}_0)\|$. Therefore, $\widehat{\xi} = O_p(\sqrt{nKq\log(Kpq)} + \sqrt{n}\|\mathbf{Z}(\widehat{\mathbf{c}}^{o} - \mathbf{c}_0)\|) = O_p(\sqrt{nKq\log(Kpq)} + \sqrt{nr\{K + (s+1)q - r\}} + n\sqrt{(s+1)qK^{-2d}}) < 1/2n\lambda$ with probability approaching 1. Thus, when $n\lambda > 2\widehat{\xi}$, and δ is chosen to be small enough, the rightmost quantity in (B.7) will be positive with probability tending to 1, thus (B.6) is proved and (30) follows. Finally, with $\widehat{f}_j^{(l)}(t) = \sum_{k=1}^{K}\widehat{c}_{jk}^{(l)}b_k(t)$, we have $\widehat{f}_j^{(l)}(t) - f_j^{(l)}(t) = \{\sum_{k=1}^{K}\widehat{c}_{jk}^{(l)}b_k(t) - \sum_{k=1}^{K}c_{0jk}^{(l)}b_k(t)\} + \{\sum_{k=1}^{K}c_{0jk}^{(l)}b_k(t) - f_j^{(l)}(t)\}$. By Condition (C5),

$$\left\|\sum_{k=1}^{K} c_{0jk}^{(l)}b_k(t) - f_j^{(l)}(t)\right\|^2 = O_p(K^{-2d}). \tag{B.8}$$

Therefore, (31) follows from (30) and (B.8).

References

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

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

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

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Chen, K., Chan, K. S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):203–221.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

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

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

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

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

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 Addisons Disease.* PhD thesis, University of Adelaide.

Gong, P., Zhang, C., Lu, Z., Huang, J. Z., and Ye, J. (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 2, pages 37– 45.

Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society. Series B*, 55:757–796.

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

Huang, J., Breheny, P., and Ma, S. (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, 27(4):481–499.

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

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

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

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

Kurtz, T. W. and Spence, M. A. (1993). Genetics of essential hypertension. *The American Journal of Medicine*, 94(1):77–84.

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

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

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013). Epigenome-wide association data implicate

dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2):142–147.

Ma, S., Carroll, R. J., Liang, H., and Xu, S. (2015). Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates. *The Annals of Statistics*, 43(5):2102–2131.

Ma, X., Xiao, L., and Wong, W. (2014). Learning regulatory programs by threshold svd regression. *Proceedings of the National Academy of Sciences*, 111:15675–15680.

Ma, Z., Ma, Z., and Sun, T. (2016). Adaptive estimation in two-way sparse reduced-rank regression. *arXiv:1403.1922v2*.

Magnus, J. R., Neudecker, H., et al. (1995). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.

McElroy, J. J. (2013). *Genetics of spontaneous idiopathic preterm birth: exploration of maternal and fetal genomes*. PhD thesis, Vanderbilt University.

Pausova, Z., Tremblay, J., and Hamet, P. (1999). Gene-environment interactions in hypertension. *Current Hypertension Reports*, 1(1):42–50.

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

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

Storlie, C., Bondell, H., Reich, B., and Zhang, H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21:679–705.

Szarek, S. J. (1982). Nets of grassmann manifold and orthogonal group. In *Proceedings of Research Workshop on Banach Space Theory*, pages 169–185.

Tabery, J. (2007). Biometric and developmental gene-environment interactions: Looking back, moving forward. *Development and Psychopathology*, 19(4):961–976.

Taylor, J. Y., Sun, Y. V., Hunt, S. C., and Kardia, S. L. (2010). Gene–environment interaction for hypertension among african american women across generations. *Biological Research for Nursing*, 12(2):149–155.

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

Wang, Y., Yin, W., and Zeng, J. (2015). Global convergence of admm in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*.

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

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

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

Yun, S., Tseng, P., and Toh, K.-C. (2011). A block coordinate gradient descent method for regularized convex separable optimization and covariance selection. *Mathematical programming*, 129(2):331–355.

Table 1: The response variables for the FHS data.

variable name	description
weight	weight (to nearest pound)
height	height (in inches to next lower 1/4 inch)
bi.deltoid.girth	bi-deltoid girth (inches with 2 decimals)
right.arm.girth.upper	right arm girth-upper third (inches with 2 decimals)
waist.girth	waist girth (inches with 2 decimals)
hip.girth	hip girth (inches with 2 decimals)
thigh.girth	thigh girth (inches with 2 decimals)
systolic.blood.pressure	systolic blood pressure-nurse
diastolic.blood.pressure	diastolic blood pressure-nurse
phy.sys.bp.1st.read	physician systolic pressure 1st reading
phy.dia.bp.1st.read	physician diastolic pressure 1st reading
phy.sys.bp.2nd.read	physician systolic pressure 2nd reading
phy.dia.bp.2nd.read	physician diastolic pressure 2nd reading
ventricular.rate	ventricular rate (per minute)
qrs.angle	qrs angle

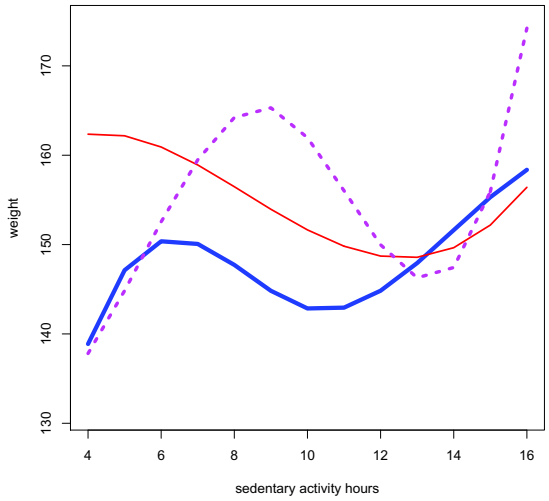

Figure 1: Plots of the estimated mean curves of weight by cubic splines against hours of sedentary activity per day for the three genotype categories *AA* (thick line), *Aa* (dashed line), and *aa* (thin line) of the SNP ss66101769 from the Framingham Heart Study (FHS), where *A* is the minor allele.

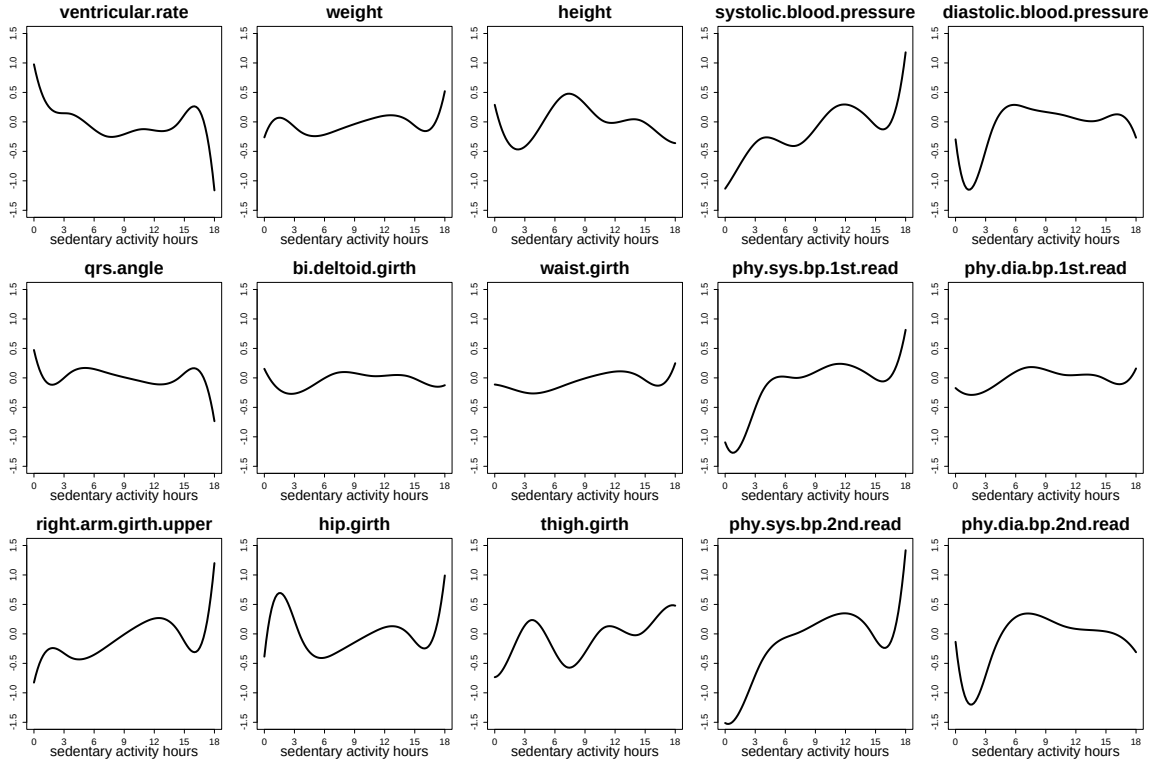

Figure 2: Estimated varying coefficients of the Biologically confirmed SNP *rs*4896044.