Graduate School of Economics, Hitotsubashi University
Discussion Paper Series No. 2018-04

# The de-biased group Lasso estimation for varying coefficient models

**Toshio HONDA**

First version：November 2018
This version：August 2019

# The de-biased group Lasso estimation for varying coefficient models

**Toshio Honda**

**Abstract**

There has been a lot of attention on the de-biased or de-sparsified Lasso since it was proposed in 2014. The Lasso is very useful in variable selection and obtaining initial estimators for other methods in high-dimensional settings. However, it is well-known that the Lasso produces biased estimators. Therefore several authors simultaneously proposed the de-biased Lasso to fix this drawback and carry out statistical inferences based on the de-biased Lasso estimators. The de-biased Lasso procedures need desirable estimators of high-dimensional precision matrices for bias correction. Thus the research is almost limited to linear regression models with some restrictive assumptions, generalized linear models with stringent assumptions and the like. To our knowledge, there are a few papers on linear regression models with group structure, but no result on structured nonparametric regression models such as varying coefficient models. In this paper, we apply the de-biased group Lasso to varying coefficient models and closely examine the theoretical properties and the effects of approximation errors involved in nonparametric regression. Some simulation results are also presented.

**Keywords**: high-dimensional data; B-spline; varying coefficient models; group Lasso; bias correction.

## 1 Introduction

We consider the following high-dimensional varying coefficient model :

$$Y_i = \sum_{j=1}^{p} g_j(Z_i)X_{i,j} + \epsilon_i, \tag{1}$$

where $(Y_i, \underline{X}_i, Z_i)$, $i = 1, \dots, n$, are i.i.d. observations, $Y_i$ is a dependent variable, $\underline{X}_i = (X_{i,1}, \dots, X_{i,p})^T \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}$ are random covariates, and an unobserved error $\epsilon_i$ follows the normal distribution with mean zero and variance $\sigma_\epsilon^2$ independently of $(\underline{X}_i, Z_i)$. Note that $a^T$ is the transpose of a vector or matrix $a$. In (1), $Z_i$ is a key variable sometimes called an index variable and $X_{i,1}$ satisfies $X_{i,1} \equiv 1$. Besides, $Z_i$ takes values on $[0, 1]$ and $g_j(Z_i)$ $j = 1, \dots, p$, are

Toshio Honda is Professor, Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan (Email: t.honda@r.hit-u.ac.jp). This research is supported by JSPS KAKENHI Grant Number JP 16K05268.

unknown smooth functions on $[0, 1]$ to be specified later in Section 3. The varying coefficient model is one of the most popular structured nonparametric regression models. For example, see [11] for an excellent review on varying coefficient models. Such structured nonparametric regression models alleviate the curse of dimensionality, but they allow much more flexibility in modelling and data analysis than linear regression models.

Nowadays a lot of high-dimensional datasets are available because of rapid advances in data collecting technology and it is inevitable to apply structured nonparametric regression models to such kinds of high-dimensional datasets for more flexible data analysis. In this paper, we take $p = O(n^{c_p})$ for some positive constant $c_p$ and this excludes ultra-high dimensional cases. This is because the technical conditions and the proofs are complicated and we give priority to readability. In practice we have to pay some cost for nonparametric estimation of coefficient functions and have some difficulty dealing with ultra-high dimensional cases. Note that the actual dimension is $pL$, where $L$ is the dimension of the spline basis.

In high-dimensional settings, even if $p$ is very large compared to the sample size $n$, the number of active or relevant covariates are much smaller than $p$ and we need some variable selection procedures for high-dimensional datasets like the Lasso(e.g.[26] and [1]), the SCAD(e.g.[7]), feature screening procedures based on marginal models or some index between the dependent variable and individual covariates(e.g.[9]), and forward variable selection procedures(e.g.[30] and [17]). [21] is an excellent review paper of feature screening procedures. The adaptive Lasso and the group Lasso are important variants of the Lasso. For example, see [35], [33], [22]. There are too many papers on high-dimensional issues to mention and we just name a few books for recent developments, [3], [13], and [27].

Several authors considered ultra-high dimensional or high-dimensional varying coefficient models by employing the group Lasso(e.g.[31]), the group SCAD(e.g.[5]), feature screening procedures based on marginal models and so on (e.g.[8] and [20]), and forward variable selection procedures(e.g.[6]). In [14] and [15], the authors considered Cox regression models with high-dimensional varying coefficient structures.

The Lasso is very useful in variable selection and obtaining initial estimators for other methods like the SCAD in high-dimensional settings. However, it is well-known that the Lasso is not necessarily selection consistent and produces biased estimators. We need some suitable initial estimators or screening procedures to reduce the number of covariates when we implement the SCAD. Screening procedures are based on marginal models or some index between $Y_i$ and individual covariates. And the procedures crucially depend on assumptions like the one that marginal models reflect the true model faithfully. When we need some reliable estimates maintaining the original high dimensionality, these procedures may not be very useful. The SCAD has the nice oracle property, but it gives no information about removed or unselected covariates. When a covariate of interest is not selected, we have no information other than being not selected. On the other hand, the de-biased Lasso gives some useful information such as p-values. The SCAD selects covariates and set the coefficient to be 0 if the covariate is not selected. Statistical inference under the original model is impossible for the SCAD.

Several authors([34], [18], and [28]) simultaneously proposed the de-biased Lasso to fix the fore-mentioned drawbacks of the Lasso and the SCAD. It is also called the de-sparsified Lasso. We can carry out statistical inferences based on the de-biased Lasso estimators while maintaining the high dimensionality and get information about all the covariates of the original high-dimensional model. The de-biased Lasso procedures need desirable estimators of high-dimensional precision matrices for bias correction. Thus the research is almost limited to linear regression models with some restrictive assumptions, generalized linear models with stringent assumptions, and the like. To our knowledge, there are a few papers on linear regression models with group structure(e.g. [23], [25]). The authors of these papers derived interesting and useful results. But we have found no result on structured nonparametric regression models such as varying coefficient models. Besides their assumptions on covariate variables cannot cover our setup since we have to deal with $\boldsymbol{W}$ defined in (4) and our design matrix $\boldsymbol{W}$ has a special structure due to the B-spline basis and $\{Z_i\}$.

We have to examine the properties carefully by carrying out conditional arguments on $\{Z_i\}$ and using the properties of the B-spline basis. We also have to take care of approximation errors to true coefficient functions. Our purpose is to estimate coefficient functions and different from that of [23] and [25] does not deal with random design cases. Both of them consider only linear models. In this paper, we apply the de-biased group Lasso to varying coefficient models and closely examine the theoretical properties of estimated coefficients and the effects of approximation errors involved in nonparametric regression.

This paper is organized as follows. In Section 2, we describe the de-biased group Lasso procedure for varying coefficient models. Then we present our assumptions and main theoretical results in Section 3. Simulation study results are presented in Section 4. The results suggest that the proposed de-biased group Lasso will work well. Additional numerical results are given in the Supplement. We prove the main theoretical results in Section 5. The technical proofs are also relegated to the supplement.

We end this section with some notation used throughout the paper.

In this paper, we write $A := B$ when we define $A$ by $B$. $C, C_1, C_2, \dots$, are generic positive constants and their values may change from line to line. Note that $a_n \sim b_n$ means $C_1 < a_n/b_n < C_2$ and that $a \vee b$ and $a \wedge b$ stand for the maximum and the minimum of $a$ and $b$, respectively.

In the theory of the group Lasso, index sets often appear and $\overline{S}$ and $|S|$ stand for the complement and the number of the elements of an index set $\mathcal{S} \subset \{1, \dots, p\}$, respectively. When we have two random vectors $U$ and $V$, $U|V$ stands for the conditional distribution of $U$ on $V$. And $\mathrm{N}(\mu, \sigma^2)$ means the normal distribution with mean $\mu$ and variance $\sigma^2$ and we write $U \sim \mathrm{N}(\mu, \sigma^2)$ when $U$ follows the normal distribution with mean $\mu$ and variance $\sigma^2$. Convergence in distribution is denoted by $\xrightarrow{d}$.

For a vector $a$, $||a||$ is the Euclidean norm and $||g||_2$ and $||g||_\infty$ stand for the $L_2$ and sup norms of a function $g$ on the unit interval, respectively. We denote the maximum and minimum eigenvalues of a symmetric matrix $A$ by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. For a matrix $A$, $||A||_F$ and $\rho(A)$ stand for the Frobenius and spectral norms, respectively. We write $(A)_{s,t}$ for the $(s, t)$ element of a matrix $A$ and $I_k$ is the $k$-dimensional identity matrix.

## 2 The de-biased group Lasso estimator

In this section, we define the de-biased group Lasso estimator $\widehat{\boldsymbol{b}}$ from the group Lasso estimator $\widehat{\boldsymbol{\beta}}$. Then we need some desirable estimator of the precision matrix of $\Sigma$ in Assumption S1 below and we denote the estimator by $\widehat{\Theta}$. We present $\widehat{\Theta}$ after we define $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$.

• **Regression spline model** : First we explain our regression spline model for (1). We denote the $L$-dimensional equispaced B-spline basis on $[0, 1]$ by $B(z) = (B_1(z), \dots, B_L(z))^T$ with $\sum_{k=1}^L B_k(z) \equiv \sqrt{L}$, not 1. We employ a quadratic or smoother basis here. The conditions on $L$ and coefficient functions are given Section 3, e.g. in Assumptions G and L.

By choosing a suitable $\beta_{0j} \in \mathbb{R}^L$, we can approximate $g_j(z)$ by $B^T(z)\beta_{0j}$ as

$$g_j(z) = B^T(z)\beta_{0j} + r_{zj}(z),$$

where $r_{zj}(z)$ is a small approximation error. Then (1) is rewritten as

$$Y_i = \sum_{j=1}^{p} X_{i,j} B^T(Z_i)\beta_{0j} + r_i + \epsilon_i, \tag{2}$$

where $r_i = \sum_{j=1}^p (g(Z_i) - B^T(Z_i)\beta_{0j})X_{i,j}$. Note that we take $\beta_{0j} = 0 \in \mathbb{R}^L$ if $g_j(z) \equiv 0$.

Now we define new $pL$-dimensional covariate vectors and the $n \times (pL)$ design matrix for the regression spline model as

$$\underline{W}_i := \underline{X}_i \otimes B(Z_i) = (X_{i,1}B^T(Z_i), \dots, X_{i,p}B^T(Z_i))^T \in \mathbb{R}^{pL}, \tag{3}$$

where $\otimes$ is the Kronecker product, and

$$\boldsymbol{W} := \begin{pmatrix} \underline{W}_1^T \\ \vdots \\ \underline{W}_n^T \end{pmatrix} = (W_1, \dots, W_p), \tag{4}$$

where $\boldsymbol{W}$ is an $n \times (pL)$ matrix and $W_j$ is an $n \times L$ matrix. Note that we have $n$ i.i.d. $\underline{W}_i \in \mathbb{R}^{pL}$. We write

$$W_j = (W_j^{(1)}, \dots, W_j^{(L)}) \quad \text{and} \quad W_j^{(l)} = \begin{pmatrix} W_{1,j}^{(l)} \\ \vdots \\ W_{n,j}^{(l)} \end{pmatrix} \in \mathbb{R}^n \text{ for } l = 1, \dots, L.$$

Note that $W_j$ is a covariate matrix for $g_j(Z_i)X_{i,j}$ and that $W_{i,j}^{(l)} = X_{i,j}B_l(Z_i)$ is an element of $\boldsymbol{W}$.

By using the above notation, we can represent $n$ observations in a matrix form :

$$Y = \sum_{j=1}^{p} W_j\beta_{0j} + r + \epsilon = \boldsymbol{W}\boldsymbol{\beta}_0 + r + \epsilon, \quad \text{where} \quad Y_i = \underline{W}_i^T\boldsymbol{\beta}_0 + r_i + \epsilon_i, \tag{5}$$

$Y = (Y_1, \dots, Y_n)^T$, $r = (r_1, \dots, r_n)^T$, $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$, and $\boldsymbol{\beta}_0 = (\beta_{01}^T, \dots, \beta_{0p}^T)^T \in \mathbb{R}^{pL}$.

We state a standard assumption on the design matrix $\boldsymbol{W}$. This is assumed throughout this paper.

**Assumption S1**

$$\Sigma := \mathrm{E}(\underline{W}_i\underline{W}_i^T) \quad \text{and} \quad \lambda_{\min}(\Sigma) > C_1$$

for some positive constant $C_1$. Note that $\Sigma$ is a $(pL) \times (pL)$ matrix.

Note that $\Sigma = n^{-1}\mathrm{E}(\boldsymbol{W}^T\boldsymbol{W})$ and we usually denote the inverse of $\Sigma$ by $\Theta$, not $\Sigma^{-1}$, as in the literature on high-dimensional precision matrices. The sample version of $\Sigma$ is $\widehat{\Sigma} := n^{-1}\boldsymbol{W}^T\boldsymbol{W}$. When $pL$ is larger than $n$, we cannot define the inverse of $\widehat{\Sigma}$. Therefore we need a reliable substitute of the inverse of $\widehat{\Sigma}$ in high-dimensional setups and we denote our estimator of the inverse $\Theta$ by $\widehat{\Theta}$. We define an $n \times (p-1)L$ matrix $\boldsymbol{W}_{-j}$ by removing $W_j$ from $\boldsymbol{W} = (W_1, \dots, W_p)$. We consider regression of $W_j$ to $\boldsymbol{W}_{-j}$ when we construct our $\widehat{\Theta}$.

- **Group Lasso estimator $\widehat{\boldsymbol{\beta}}$** : We define the group Lasso estimator $\widehat{\boldsymbol{\beta}}$ for (2) and (5) :

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1^T, \dots, \widehat{\beta}_p^T)^T := \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{pL}} \Big\{ \frac{1}{n} \|Y - \boldsymbol{W}\boldsymbol{\beta}\|^2 + 2\lambda_0 \mathrm{P}_1(\boldsymbol{\beta}) \Big\}, \tag{6}$$

where $\boldsymbol{\beta} = (\beta_1^T, \dots, \beta_p^T)^T$ with $\beta_j \in \mathbb{R}^L$ for $j = 1, \dots, p$, $\lambda_0$ is a suitably chosen tuning parameter, and $\mathrm{P}_1(\boldsymbol{\beta}) := \sum_{j=1}^{p} \|\beta_j\|$. We also use this $\mathrm{P}_1(\cdot)$ for vectors of smaller dimension. We describe the properties of this group Lasso estimator in Proposition 1 for completeness although the proposition is almost known.

The first order condition of the optimality of $\widehat{\boldsymbol{\beta}}$ yields

$$-\frac{1}{n}\boldsymbol{W}^T(Y - \boldsymbol{W}\widehat{\boldsymbol{\beta}}) + \lambda_0\boldsymbol{\kappa}_0 = 0 \in \mathbb{R}^{pL}, \tag{7}$$

where $\boldsymbol{\kappa}_0 = (\kappa_{0,1}^T, \dots, \kappa_{0,p}^T)^T$ with $\kappa_{0,j} \in \mathbb{R}^L$ for $j = 1, \dots, p$, $\|\kappa_{0,j}\| \le 1$ for $j = 1, \dots, p$, and $\kappa_{0,j} = \widehat{\beta}_j / \|\widehat{\beta}_j\|$ if $\|\widehat{\beta}_j\| \neq 0$.

• **De-biased group Lasso estimator $\widehat{\boldsymbol{b}}$** : This $\widehat{\boldsymbol{\beta}}$ is a biased estimator due to the $L_1$ penalty as we mentioned in Section 1. Thus by constructing $\widehat{\Theta}$ such that $\widehat{\Theta\Sigma}$ is sufficiently close to $I_{pL}$, we define our de-biased group Lasso estimator $\widehat{\boldsymbol{b}} = (\widehat{b}_1^T, \dots, \widehat{b}_p^T)^T \in \mathbb{R}^{pL}$ with $\widehat{b}_j \in \mathbb{R}^L$ for $j = 1, \dots, p$ for the varying coefficient model (1) and (5) as

$$\begin{aligned}\widehat{\boldsymbol{b}} &:= \widehat{\boldsymbol{\beta}} + \widehat{\Theta}\lambda_0\kappa_0 = \widehat{\boldsymbol{\beta}} + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T(Y - \boldsymbol{W}\widehat{\boldsymbol{\beta}}) \qquad (8)\\ &= \widehat{\boldsymbol{\beta}} + \widehat{\Theta\Sigma}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}) + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T(r + \epsilon)\\ &= \boldsymbol{\beta}_0 + \frac{1}{n}(\widehat{\Theta\Sigma} - I_{pL})(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}) + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T(r + \epsilon)\\ &= \boldsymbol{\beta}_0 + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T\epsilon - \Delta_1 + \Delta_2,\end{aligned}$$

where we used (7) in the first line,

$$\begin{aligned}\Delta_1 &= (\Delta_{1,1}^T, \dots, \Delta_{1,p}^T)^T := \frac{1}{n}(\widehat{\Theta\Sigma} - I_{pL})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \in \mathbb{R}^{pL},\\ \Delta_2 &= (\Delta_{2,1}^T, \dots, \Delta_{2,p}^T)^T := \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T r \in \mathbb{R}^{pL},\end{aligned}$$

$$\Delta_1 = \begin{pmatrix}\Delta_{1,1}\\ \vdots\\ \Delta_{1,p}\end{pmatrix} := \frac{1}{n}(\widehat{\Theta\Sigma} - I_{pL})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \in \mathbb{R}^{pL}, \quad \Delta_2 = \begin{pmatrix}\Delta_{2,1}\\ \vdots\\ \Delta_{2,p}\end{pmatrix} := \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T r \in \mathbb{R}^{pL},$$

and $\Delta_{1,j} \in \mathbb{R}^L$ and $\Delta_{2,j} \in \mathbb{R}^L$ for $j = 1, \dots, p$. We will prove that $\Delta_1$ and $\Delta_2$ are negligible compared to $n^{-1}\widehat{\Theta}\boldsymbol{W}^T\epsilon$ in Propositions 3 and 4, respectively and closely examine $n^{-1}\widehat{\Theta}\boldsymbol{W}^T\epsilon$ in Proposition 5 in Section 3.

The evaluation of $\Delta_2$ requires more smoothness of the coefficient functions $g_j(z)$ than usual as in Assumption G in Section 3. This is because it is difficult to evaluate the effects of approximation errors while maintaining high-dimensionality as shown in the proof of Proposition 3. Any model may have some kind of approximation error and it is very important to examine such effects in the de-biased Lasso method closely. If we are interested in only some of $X_{i,1}, \dots, X_{i,p}$, not all of them, we do not have to compute the whole $\widehat{\boldsymbol{b}}$ and should concentrate on only the corresponding blocks.

• **Construction of $\widehat{\Theta}$** : At the end of this section, we construct $\widehat{\Theta}$ by employing the group Lasso and adapting the idea in [28] to the current group structure. Note that our construction is different from those of [23] and [25] and that we can exploit just the standard R package for the Lasso for computation. We also describe some idea of how to $\widehat{\Theta}$ in (9)-(11) after the notation.

We need some more notation before we present our $\widehat{\Theta}$. Hereafter, we write $a^{\otimes 2} := aa^T$ for a vector $a$. We define an $L \times L$ matrix $\Sigma_{j,k}$, an $L \times (p-1)L$ matrix $\Sigma_{j,-j}$, a $(p-1)L \times L$ matrix

$\Sigma_{-j,j}$, and a $(p-1)L \times (p-1)L$ matrix $\Sigma_{-j,-j}$ :

$$\Sigma_{j,k} := \mathrm{E}\{X_{1,j}X_{1,k}B^{\otimes 2}(Z_1)\} = \frac{1}{n}\mathrm{E}(W_j^T W_k)$$
$$\Sigma_{j,-j} := \mathrm{E}[\{X_{1,j}(X_{1,1},\dots,X_{1,j-1},X_{1,j+1},\dots,X_{1,p})\} \otimes B^{\otimes 2}(Z_1)] = \frac{1}{n}\mathrm{E}(W_j^T \boldsymbol{W}_{-j})$$
$$\Sigma_{-j,-j} := \mathrm{E}[\{(X_{1,1},\dots,X_{1,j-1},X_{1,j+1},\dots,X_{1,p})^T\}^{\otimes 2} \otimes B^{\otimes 2}(Z_1)] = \frac{1}{n}\mathrm{E}(\boldsymbol{W}_{-j}^T \boldsymbol{W}_{-j})$$

and $\Sigma_{-j,j} := \Sigma_{j,-j}^T$. Note that they can be defined also from $\Sigma$ as its submatrices. Furthermore we define a $(p-1)L \times L$ matrix $\Gamma_j$ as $\Gamma_j := \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$ and write $\Gamma_j = (\gamma_j^{(1)},\dots,\gamma_j^{(L)})$, where $\gamma_j^{(l)} \in \mathbb{R}^{(p-1)L}$ for $l = 1,\dots,L$. We need to estimate this $\Gamma_j$ to define $\widehat{\Theta}$. In this paper, we estimate $\Gamma_j = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j} = (\gamma_l^{(1)},\dots,\gamma_l^{(L)})$ columnwise by employing the group Lasso differently from [25]. See Remark 1 at the end of this section.

To present an idea on the construction of $\widehat{\Theta}$, we give some insightful expressions such as (10)-(12). Then we define an $n \times L$ matrix $E_j$ and its columns $\eta_j^{(l)} \in \mathbb{R}^n$, $j = 1,\dots,L$, as

$$E_j = (\eta_j^{(1)},\dots,\eta_j^{(L)}) := W_j - \boldsymbol{W}_{-j}\Gamma_j. \tag{9}$$

Since $\Sigma_{-j,j} - \Sigma_{-j,-j}\Gamma_j = n^{-1}\mathrm{E}(\boldsymbol{W}_{-j}^T E_j) = 0$, we have

$$\begin{aligned}
\frac{1}{n}\mathrm{E}(\boldsymbol{W}^T E_1) &= \frac{1}{n}\mathrm{E}\{\boldsymbol{W}^T(W_1 - \boldsymbol{W}_{-1}\Gamma_1)\} = (\Theta_{1,1}^{-1},0,0,\dots,0)^T \\
&\cdots\cdots \\
\frac{1}{n}\mathrm{E}(\boldsymbol{W}^T E_j) &= \frac{1}{n}\mathrm{E}\{\boldsymbol{W}^T(W_j - \boldsymbol{W}_{-j}\Gamma_j)\} = (0,\Theta_{j,j}^{-1},0,\dots,0)^T \qquad (10) \\
&\cdots\cdots \\
\frac{1}{n}\mathrm{E}(\boldsymbol{W}^T E_p) &= \frac{1}{n}\mathrm{E}\{\boldsymbol{W}^T(W_p - \boldsymbol{W}_{-p}\Gamma_p)\} = (0,\dots,0,0,\Theta_{p,p}^{-1})^T,
\end{aligned}$$

where symmetric $L \times L$ matrices $\Theta_{j,j}$ will be defined shortly. The above equations imply

$$\frac{1}{n}\mathrm{E}\{\boldsymbol{W}^T(E_1,\dots,E_p)\}\begin{pmatrix} \Theta_{1,1}^{-1} & 0 & 0 & \cdots & 0 \\ 0 & \Theta_{2,2}^{-1} & 0 & \cdots & 0 \\ 0 & 0 & \Theta_{3,3}^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Theta_{p,p}^{-1} \end{pmatrix}^{-1} = I_{pL}. \tag{11}$$

Recalling that $n^{-1}\mathrm{E}(\boldsymbol{W}^T\boldsymbol{W}) = \Sigma$ and (9), we define $\widehat{\Theta}$ by employing the sample version of the LHS of (11). Thus we need to estimate $\Gamma_j$, $j = 1,\dots,p$. See also (19) below.

Let $\Theta_{j,k}$ be an $L \times L$ submatrix of $\Theta$ exactly as $\Sigma_{j,k}$ is a submatrix of $\Sigma$. Then we have

$$\Theta_{j,j}^{-1} = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} = \frac{1}{n}\mathrm{E}(E_j^T E_j) = \frac{1}{n}\mathrm{E}(W_j^T E_j). \tag{12}$$

We explain how we estimate $\Gamma_j$. Looking at (9) and $n^{-1}\mathrm{E}(\boldsymbol{W}_{-j}^T E_j) = 0$ columnwise, we have

$$\eta_j^{(l)} = W_j^{(l)} - \boldsymbol{W}_{-j}\gamma_j^{(l)} \in \mathbb{R}^n, \quad l = 1, \dots, L \text{ and } j = 1, \dots, p,$$

and then we estimate $\Gamma_j = (\gamma_l^{(1)}, \dots, \gamma_l^{(L)})$ columnwise by employing the group Lasso:

$$\widehat{\gamma}_j^{(l)} = (\widehat{\gamma}_{j,1}^{(l)T}, \dots, \widehat{\gamma}_{j,j-1}^{(l)T}, \widehat{\gamma}_{j,j+1}^{(l)T}, \dots, \widehat{\gamma}_{j,p}^{(l)T})^T := \operatorname*{argmin}_{\gamma \in \mathbb{R}^{(p-1)L}} \Big\{ \frac{1}{n} \|W_j^{(l)} - \boldsymbol{W}_{-j}\gamma\|^2 + 2\lambda_j^{(l)} \mathrm{P}_1(\gamma) \Big\}, \tag{13}$$

where $\mathrm{P}_1(\gamma)$ is defined as in (6), $\widehat{\gamma}_{j,k}^{(l)} \in \mathbb{R}^L$ for $k \neq j$, $\gamma = (\gamma_1^T, \dots, \gamma_{j-1}^T, \gamma_{j+1}^T, \dots, \gamma_p^T)^T$ with $\gamma_k \in \mathbb{R}^L$ for $k \neq j$, and $\lambda_j^{(l)}$ is a suitably chosen tuning parameter. We deal with the theoretical properties of $\widehat{\gamma}_j^{(l)}$ in Proposition 2 in Section 3.

As in (7), we have

$$-\frac{1}{n}\boldsymbol{W}_{-j}^T(W_j^{(l)} - \boldsymbol{W}_{-j}\widehat{\gamma}_j^{(l)}) + \lambda_j^{(l)}\boldsymbol{\kappa}_j^{(l)} = 0 \in \mathbb{R}^{(p-1)L}, \tag{14}$$

where $\boldsymbol{\kappa}_j^{(l)} = (\kappa_{j,1}^{(l)T}, \dots, \kappa_{j,j-1}^{(l)T}, \kappa_{j,j+1}^{(l)T}, \dots, \kappa_{j,p}^{(l)T})^T$ with $\kappa_{j,k}^{(l)} \in \mathbb{R}^L$ for $k \neq j$, $\|\kappa_{j,k}^{(l)}\| \leq 1$ for $k \neq j$, and $\kappa_{j,k}^{(l)} = \widehat{\gamma}_{j,k}^{(l)}/\|\widehat{\gamma}_{j,k}^{(l)}\|$ if $\|\widehat{\gamma}_{j,k}^{(l)}\| \neq 0$.

Collecting $\widehat{\gamma}_j^{(l)}$, $\kappa_{j,k}^{(l)}$, and regression residuals into matrices, respectively, we define $(p-1)L \times L$ matrices $\widehat{\Gamma}_j$ and $K_j$ and an $n \times L$ matrix $\widehat{E}_j$ as

$$\widehat{\Gamma}_j := (\widehat{\gamma}_j^{(1)}, \dots, \widehat{\gamma}_j^{(L)}), \quad K_j := (\boldsymbol{\kappa}_j^{(1)}, \dots, \boldsymbol{\kappa}_j^{(L)}), \quad \text{and} \quad \widehat{E}_j := W_j - \boldsymbol{W}_{-j}\widehat{\Gamma}_j \tag{15}$$

and write

$$\begin{aligned} &\widehat{\Gamma}_j = (\widehat{\Gamma}_{j,1}^T, \dots, \widehat{\Gamma}_{j,j-1}^T, \widehat{\Gamma}_{j,j+1}^T, \dots, \widehat{\Gamma}_{j,p}^T)^T, \quad K_j = (K_{j,1}^T, \dots, K_{j,j-1}^T, K_{j,j+1}^T, \dots, K_{j,p}^T)^T, \quad \text{and} \\ &\widehat{E}_j = \boldsymbol{W}(-\widehat{\Gamma}_{j,1}^T, \dots, -\widehat{\Gamma}_{j,j-1}^T, I_L, -\widehat{\Gamma}_{j,j+1}^T, \dots, -\widehat{\Gamma}_{j,p}^T)^T, \end{aligned} \tag{16}$$

where $\widehat{\Gamma}_{j,k}$ $(k \neq j)$ and $K_{j,k}$ $(k \neq j)$ are $L \times L$ matrices. Then by (14), we have

$$\frac{1}{n}\boldsymbol{W}_{-j}^T\widehat{E}_j = K_j\Lambda_j, \tag{17}$$

where $\Lambda_j = \mathrm{diag}(\lambda_j^{(1)}, \dots, \lambda_j^{(L)})$. The elements of the RHS of (17) will be small because of the properties of the group Lasso. Recall that $n^{-1}\mathrm{E}(\boldsymbol{W}_{-j}^T E_j) = 0$.

We are ready to define $\widehat{\Theta}$ by adapting the idea of [28] to the current setup. Let $T_j^2$ be our estimator of $\Theta_{j,j}^{-1}$ and defined later. See also (9), (11), and (16).

$$\widehat{\Theta}^T := \begin{pmatrix} I_L & -\widehat{\Gamma}_{2,1} & -\widehat{\Gamma}_{3,1} & \cdots & -\widehat{\Gamma}_{p,1} \\ -\widehat{\Gamma}_{1,2} & I_L & -\widehat{\Gamma}_{3,2} & \cdots & -\widehat{\Gamma}_{p,2} \\ -\widehat{\Gamma}_{1,3} & -\widehat{\Gamma}_{2,3} & I_L & \cdots & -\widehat{\Gamma}_{p,3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\widehat{\Gamma}_{1,p} & -\widehat{\Gamma}_{2,p} & -\widehat{\Gamma}_{3,p} & \cdots & I_L \end{pmatrix} \begin{pmatrix} T_1^2 & 0 & 0 & \cdots & 0 \\ 0 & T_2^2 & 0 & \cdots & 0 \\ 0 & 0 & T_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & T_p^2 \end{pmatrix}^{-1}. \tag{18}$$

Hereafter the second matrix on the RHS will be denoted by $\mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2})$. Note that $T_j^{-2}$ stands for the inverse of $T_j^2$ and is an estimator of $\Theta_{j,j}$.

(16)-(18) give the following equations if we take $T_j^2 := \frac{1}{n} W_j^T \widehat{E}_j$. Compare (11) and (19), too.

$$\widehat{\Sigma}\widehat{\Theta}^T = \frac{1}{n}\boldsymbol{W}^T(\boldsymbol{W}\widehat{\Theta}^T) = \frac{1}{n}\boldsymbol{W}^T(\widehat{E}_1, \ldots, \widehat{E}_p)\mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2}) \tag{19}$$

$$= \begin{pmatrix} T_1^2 & K_{2,1}\Lambda_2 & K_{3,1}\Lambda_3 & \cdots & K_{p,1}\Lambda_p \\ K_{1,2}\Lambda_1 & T_2^2 & K_{3,2}\Lambda_3 & \cdots & K_{p,2}\Lambda_p \\ K_{1,3}\Lambda_1 & K_{2,3}\Lambda_2 & T_3^2 & \cdots & K_{p,3}\Lambda_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{1,p}\Lambda_1 & K_{2,p}\Lambda_2 & K_{3,p}\Lambda_3 & \cdots & T_p^2 \end{pmatrix} \mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2})$$

and

$$\widehat{\Sigma}\widehat{\Theta}^T - I_{pL} = \begin{pmatrix} 0 & K_{2,1}\Lambda_2 T_2^{-2} & K_{3,1}\Lambda_3 T_3^{-2} & \cdots & K_{p,1}\Lambda_p T_p^{-2} \\ K_{1,2}\Lambda_1 T_1^{-2} & 0 & K_{3,2}\Lambda_3 T_3^{-2} & \cdots & K_{p,2}\Lambda_p T_p^{-2} \\ K_{1,3}\Lambda_1 T_1^{-2} & K_{2,3}\Lambda_2 T_2^{-2} & 0 & \cdots & K_{p,3}\Lambda_p T_p^{-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{1,p}\Lambda_1 T_1^{-2} & K_{2,p}\Lambda_2 T_2^{-2} & K_{3,p}\Lambda_3 T_3^{-2} & \cdots & 0 \end{pmatrix}. \tag{20}$$

The elements of the off-diagonal blocks will be small due to the properties of the group Lasso in (13).

Taking the transpose of (20), we obtain

$$\widehat{\Theta}\widehat{\Sigma} - I_{pL} = \begin{pmatrix} 0 & T_1^{-2T}\Lambda_1 K_{1,2}^T & T_1^{-2T}\Lambda_1 K_{1,3}^T & \cdots & T_1^{-2T}\Lambda_1 K_{1,p}^T \\ T_2^{-2T}\Lambda_2 K_{2,1}^T & 0 & T_2^{-2T}\Lambda_2 K_{2,3}^T & \cdots & T_2^{-2T}\Lambda_2 K_{2,p}^T \\ T_3^{-2T}\Lambda_3 K_{3,1}^T & T_3^{-2T}\Lambda_3 K_{3,2}^T & 0 & \cdots & T_3^{-2T}\Lambda_3 K_{3,p}^T \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_p^{-2T}\Lambda_p K_{p,1}^T & T_p^{-2T}\Lambda_p K_{p,2}^T & T_p^{-2T}\Lambda_p K_{p,3}^T & \cdots & 0 \end{pmatrix}. \tag{21}$$

We denote $(T_j^{-2})^T$ by $T_j^{-2T}$. We will closely examine

$$T_j^{-2T}\Lambda_j K_{j,k}^T = T_j^{-2T} \begin{pmatrix} \lambda_j^{(1)} \kappa_{j,k}^{(1)T} \\ \vdots \\ \lambda_j^{(L)} \kappa_{j,k}^{(L)T} \end{pmatrix}$$

to deal with $\Delta_1$ in Proposition 3.

Finally note that $T_j^2 = n^{-1} W_j^T \widehat{E}_j$, our estimator of $\Theta_{j,j}^{-1} = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$, satisfies $\min_{1 \le j \le p} \rho(T_j^2) > C$ with probability tending to 1 for some constant $C$ as proved in Lemma 6 in Section 5. See also (12) about this definition of $T_j^2$.

In Section 4, we chose $\lambda_0$ and $\lambda_j^{(l)}$ by cross validation. In the next section, we give theoretically proper ranges of these tuning parameters. But we have no theory for tuning parameter selection.

**Remark 1** In [25], the authors considered fixed design regression models and estimated all the columns of $\Gamma_j$ simultaneously in a single Lasso-type penalized regression. On the other hand, we estimate $\Gamma_j$ columnwise and we can apply the standard theory and also employ the standard R package to get our estimator of $\Gamma_j$. We can define another estimator of $\Theta$ just formally even if we estimate $\Gamma_j$ simultaneously. Then the properties will be different from those of this paper and we cannot apply the standard Lasso theory and R packages then.

# 3 Theoretical results

In this section, we state the standard result on the group Lasso estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}_j^{(l)}$ in Propositions 1 and 2 together with technical assumptions. Then we evaluate $\Delta_1$ and $\Delta_2$ in (8) and $\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^T$ in Propositions 3-5. Finally we state the main result on the de-biased group Lasso estimator $\widehat{\boldsymbol{b}}$ in Theorem 1. We will prove Propositions 3-5 in Section 5. Theorem 1 immediately follows from those propositions. Propositions 1 and 2 will be proved in the Supplement since we can prove them by just following the standard arguments in the Lasso literature. The proofs of all the technical lemmas will also be given in the Supplement.

• **Basic assumptions** : We describe some notation and assumptions before we present the results on the group Lasso estimators. We define the set of active covariates and the number of active covariates :

$$\mathcal{S}_0 := \{j \,|\, \|g_j\|_2 > 0\} \subset \{1, \dots, p\} \quad \text{and} \quad s_0 := |\mathcal{S}_0|. \tag{22}$$

We begin with some definitions to state basic assumptions on the properties of covariates of our varying coefficient model:

$$\underline{\widetilde{X}}_i = (X_{i,2}, \dots, X_{i,p})^T \quad \text{and} \quad \underline{\check{X}}_i = \underline{\widetilde{X}}_i - \mu_X(Z_i),$$

where $\mu_X(Z_i) = (\mu_{X,2}(Z_i), \dots, \mu_{X,p}(Z_i))^T$ is the conditional mean of $\underline{\widetilde{X}}_i$ given $Z_i$. We denote the conditional covariance matrix of $\underline{\widetilde{X}}_i$ given $Z_i$ by $\Sigma_X(Z_i)$. We define $\underline{\widetilde{X}}_i$ by removing the first constant element from $\underline{X}_i$ defined in (1).

**Assumption VC**

**(1)** $\mathrm{E}(X_{i,j}) = 0,\ j = 2, \dots, p$. Besides, $\|\mu_{X,j}\|_\infty < C_1$ for $j = 2, \dots, p$ and $C_2 < \lambda_{\min}(\Sigma_X(z)) \le \lambda_{\max}(\Sigma_X(z)) < C_3$ uniformly in $z$ on $[0, 1]$ for some positive constants $C_1$, $C_2$, and $C_3$. Recall that $\epsilon_i \sim \mathrm{N}(0, \sigma_\epsilon^2)$ and $\epsilon_i$ is independent of $(\underline{X}_i, Z_i)$.

**(2)** There is a constant $\sigma^2$ independent of $Z_i$ such that $\mathrm{E}\{\exp(\alpha^T \underline{\check{X}}_i)|Z_i\} \leq \exp(\|\alpha\|^2\sigma^2/2)$ for any $\alpha \in \mathbb{R}^{p-1}$.

**(3)** The index variable $Z_i$ has density $f_Z(z)$ satisfying $C_4 < f_Z(z) < C_5$ on $[0,1]$ for some positive constants $C_4$ and $C_5$.

The second one, the sub-Gaussian design assumption, allows us to use Bernstein's inequality. The first two assumptions may look restrictive. However, we need to construct a desirable estimator of a precision matrix and even more restrictive assumptions such as normality are imposed in [28] and [19]. Especially, the arguments in [19] crucially depend on the normality assumption on the design matrix although it has improved the previous results on the de-biased Lasso. The assumption on $\{\epsilon_i\}$ is the standard one in the literature of the de-biased Lasso. In [4], the authors developed the theory of the de-biased Lasso for linear models without normality or sub-Gaussian assumption on design matrices, but they need a restrictive assumption on the order of $p$ such as $p \ll n$ and other alternative conditions. The third one is a standard assumption for varying coefficient models.

Next we state the assumptions on coefficient functions.

**Assumption G**

**(1)** $g_j(z)$, $j = 1, \ldots, p$, are three times continuously differentiable.

**(2)** If we choose suitable $\beta_{0j} \in \mathbb{R}^L$ and $d_j$ $j = 1, \ldots, p$, the approximation error $r_{i,j}$ defined as $r_{i,j} = g_j(Z_i) - B^T(Z_i)\beta_{0j}$ satisfies

$$|r_{i,j}| < C_1 L^{-3} d_j \text{ for } i = 1, \ldots, n \text{ and } j \in \mathcal{S}_0, \quad \sum_{j \in \mathcal{S}_0} d_j < C_2, \quad \text{and} \quad \sum_{j \in \mathcal{S}_0} d_j^2 < C_3 \tag{23}$$

for some positive constants $C_1$, $C_2$, and $C_3$.

In this paper, $\sum_{k=1}^L B_k(z) \equiv \sqrt{L}$. Then we have for some positive constants $C_1$ and $C_2$ that $C_1 < \lambda_{\min}(\Omega_B) \leq \lambda_{\max}(\Omega_B) < C_2$, where $\Omega_B = \int_0^1 B(z)B^T(z)dz$. See e.g. [16] about this fact. We employ a quadratic or smoother basis. We give a remark on other spline bases in Remark 2 later in this section.

The former half of Assumption G may be a little more restrictive. However, we need this assumption to evaluate $\Delta_2$. If we take $d_j = \|g_j\|_\infty + \|g_j'\|_\infty + \|g_j''\|_\infty + \|g_j^{(3)}\|_\infty$ and some suitable $\beta_{0j}$, this $\{d_j\}$ satisfies the first one in (23). See e.g. Corollary 6.26 of [24]. This $\{d_j\}$ should satisfy the second and third ones in (23). Note that we take $\beta_{0j} = 0$ for $j \notin \mathcal{S}_0$ and that $g_j^{(3)}(z)$ is the third order derivative of $g_j(z)$.

We denote the conditional mean and variance of $L^3 \sum_{j \in \mathcal{S}_0} r_{i,j} X_{i,j}$ given $Z_i$ by $\mu_r(Z_i)$ and $\sigma_r^2(Z_i)$, respectively. Then under Assumptions VC and G, we have

$$\|\mu_r\|_\infty < C_1, \quad \text{and} \quad \|\sigma_r^2\|_\infty < C_2$$

for some positive constants $C_1$ and $C_2$. The above results, the sub-Gaussian design assumption, and the use of Bernstein's inequality imply

$$|r_i| < C_3(\log n)^{1/2}L^{-3} \tag{24}$$

uniformly in $i$ with probability tending to 1 for some positive constant $C_3$. Recall $r_i$ is defined in (2).

• **Results on $\widehat{\boldsymbol{\beta}}$** : The theoretical results on the Lasso crucially depends on the deviation condition (Lemma 1) and the restrictive eigenvalue (RE) condition or a similar one (Lemma 2). If both of the conditions are established, the standard theoretical results (Proposition 1) follow almost automatically from them.

**Lemma 1** *Suppose that Assumptions VC and G hold and that* $(L^{-3}\log n + \sqrt{n^{-1}L\log n}) \to 0$. *Then for some large constant C, we have*

$$\mathrm{P}_\infty(n^{-1}\boldsymbol{W}^T\epsilon) < C\sqrt{\frac{L\log n}{n}} \quad \text{and} \quad \mathrm{P}_\infty(n^{-1}\boldsymbol{W}^T r) < CL^{-3}\log n$$

*with probability tending to 1, where* $\mathrm{P}_\infty(\boldsymbol{v}) := \max_{1\le j\le p}\|v_j\|$ *for* $\boldsymbol{v} = (v_1^T, \dots, v_p^T)^T \in \mathbb{R}^{pL}$ *with* $v_j \in \mathbb{R}^L$ *for* $j = 1, \dots, p$.

We also use $\mathrm{P}_\infty(\cdot)$ for vectors of lower dimension as we use $\mathrm{P}_1(\cdot)$.

Some preparations are necessary to define the RE condition. For an index set $\mathcal{S} \subset \{1, \dots, p\}$ and a positive constant $m$, we define a subset of $\mathbb{R}^{pL}$ as in the literature on the Lasso :

$$\Psi(\mathcal{S}, m) := \{\boldsymbol{\beta} \in \mathbb{R}^{pL} \,|\, \mathrm{P}_1(\boldsymbol{\beta}_{\overline{\mathcal{S}}}) \le m\mathrm{P}_1(\boldsymbol{\beta}_{\mathcal{S}}) \text{ and } \boldsymbol{\beta} \neq 0\},$$

where $\boldsymbol{\beta}_{\mathcal{S}}$ consists of $\{\beta_j\}_{j\in\mathcal{S}}$, $\boldsymbol{\beta}_{\overline{\mathcal{S}}}$ consists of $\{\beta_j\}_{j\in\overline{\mathcal{S}}}$, and $\mathrm{P}_1(\cdot)$ is conformably adjusted to the dimension of the arguments. Recall $\beta_j \in \mathbb{R}^L$ in this paper. Then we define $\phi^2_\Omega(\mathcal{S}, m)$ for a non-negative $(pL) \times (pL)$ matrix $\Omega$ as

$$\phi^2_\Omega(\mathcal{S}, m) := \min_{\boldsymbol{\beta}\in\Psi(\mathcal{S},m)} \frac{\boldsymbol{\beta}^T\Omega\boldsymbol{\beta}}{\|\boldsymbol{\beta}_{\mathcal{S}}\|^2}.$$

In the theory of the Lasso, $\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, m)$ plays a crucial role and the lower bound is given in Lemma 2 below.

**Lemma 2** *Suppose that Assumptions VC and S1 hold and that* $s_0\sqrt{n^{-1}L^3\log n} \to 0$. *Then*

$$2\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3) \ge \phi^2_\Sigma(\mathcal{S}_0, 3) \ge \lambda_{\min}(\Sigma)$$

*with probability tending to 1.*

Notice that the second inequality is trivial from the definition of $\phi^2_\Sigma(\mathcal{S}_0, 3)$ and always holds.

The next result may be almost known, but we present and prove it for completeness.

**Proposition 1** *Suppose that Assumptions VC, S1, and G hold and that* $(s_0\sqrt{n^{-1}L^3\log n}) \vee (L^{-3}\log n + \sqrt{n^{-1}L\log n}) \to 0$. *Then if* $\lambda_0 = C(L^{-3}\log n + \sqrt{n^{-1}L\log n})$ *for sufficiently large* $C$*, we have with probability tending to 1,*

$$\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 \leq 18\frac{\lambda_0^2 s_0}{\phi^2_\Sigma(\mathcal{S}_0, 3)} \quad \text{and} \quad \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq 24\frac{\lambda_0 s_0}{\phi^2_\Sigma(\mathcal{S}_0, 3)}.$$

We will prove this proposition in the Supplement including the case where we have some prior knowledge on $\mathcal{S}_0$. Note that $C$ in the definition of $\lambda_0$ is from Lemma 1. We will follow the proof in [4] and we can also deal with the weighted group Lasso as in [4] with just conformable changes. Note that [4] considered the adaptive Lasso for linear regression models. We didn't present the adaptively weighted Lasso version since the notation is very complicated in the current setup of the group Lasso procedures. If an estimator has the oracle property, e.g. the SCAD estimator and a kind of suitably weighted Lasso estimators as in [10], it is not biased and we don't have to apply the de-biased procedure to those estimators. However, as we mentioned before, no statistical inference is possible while maintaining the original high-dimensionality.

- **Results on $\widehat{\gamma}^{(l)}_j$ for $\widehat{\Theta}$ :** We consider the properties of another group Lasso estimator $\widehat{\gamma}^{(l)}_j$ defined in (13). We deal with the deviation condition and the RE condition in Lemma 3 and Lemma 4, respectively. Then Proposition 2 about the group Lasso estimator $\widehat{\gamma}^{(l)}_j$ in (13) follows almost automatically from them.

We define the active index set $\mathcal{S}^{(l)}_j \subset \{1, \dots, j-1, j+1, \dots, p\}$ of $\gamma^{(l)}_j$ in almost the same way as $\mathcal{S}_0$ of $\boldsymbol{\beta}_0$ and let $s^{(l)}_j := |\mathcal{S}^{(l)}_j|$. We assume $\mathcal{S}^{(l)}_j$ is not empty since we can include some index in it even if it is actually empty.

We need some technical assumptions.

**Assumption S2**

**(1)** $\|\gamma^{(l)}_j\| \leq C_1$ uniformly in $l$ $(1 \leq l \leq L)$ and $j$ $(1 \leq j \leq p)$ for some positive constant $C_1$.

**(2)** $\lambda_{\max}(\Sigma_{j,j}) \leq C_2$ uniformly in $j$ $(1 \leq j \leq p)$ for some positive constant $C_2$.

Assumptions S1 and S2(2) imply

$$C_3 \leq \lambda_{\min}(\Theta^{-1}_{j,j}) = \frac{1}{\lambda_{\max}(\Theta_{j,j})} \leq \lambda_{\max}(\Theta^{-1}_{j,j}) = \frac{1}{\lambda_{\min}(\Theta_{j,j})} \leq C_4 \tag{25}$$

uniformly in $j$ for some positive constants $C_3$ and $C_4$.

We give some comments on the implications of Assumptions VC, S1, and S2 to consider the properties of the group Lasso estimator of $\gamma^{(l)}_j$ in (13). Then we write $\eta^{(l)}_j = (\eta^{(l)}_{1,j}, \dots, \eta^{(l)}_{n,j})^T \in$

$\mathbb{R}^n$. Since $\Sigma_{-j,j} - \Sigma_{-j,-j}\Gamma_j = 0$ and our observations are i.i.d., we have

$$\mathrm{E}(\underline{W}_{i,-j}\eta_{i,j}^{(l)}) = 0 \in \mathbb{R}^{(p-1)L}, \quad i = 1, \ldots, n, \ l = 1, \ldots, L, \text{ and } j = 1, \ldots, p, \tag{26}$$

where define $\underline{W}_{i,-j}$ by removing $X_{i,j}B(Z_i)$ from $\underline{W}_i$ and we have $\boldsymbol{W}_{-j} = (\underline{W}_{1,-j}, \ldots, \underline{W}_{n,-j})^T$.

We denote the conditional mean and variance of $\eta_{i,j}^{(l)}$ given $Z_i$ by $\mu_{\eta,j}^{(l)}(Z_i)$ and $\sigma_{\eta,j}^{(l)2}(Z_i)$, respectively. Under Assumption S2(2), we have

$$\mathrm{E}\{\eta_{i,j}^{(l)2}\} = \mathrm{E}[\{\mu_{\eta,j}^{(l)}(Z_i)\}^2 + \sigma_{\eta,j}^{(l)2}(Z_i)] = O(1) \tag{27}$$

uniformly in $l$ $(1 \le l \le L)$ and $j$ $(1 \le j \le p)$. Besides, Assumptions VC and S2(1), and the properties of the B-spline basis suggest

$$\|\sigma_{\eta,j}^{(l)2}\|_\infty = O(L) \tag{28}$$

uniformly in $l$ and $j$. Assumption S1 is closely related to Assumption S2(1) since $\Gamma_j = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$.

We need an assumption on $\mu_{\eta,j}^{(l)}(z)$ similar to (28) to deal with the deviation condition. We give a comment on this assumption in Remark 3 at the end of this section.

**Assumption E** Under Assumption VC, we have uniformly in $l$ $(1 \le l \le L)$ and $j$ $(1 \le j \le p)$ ,

$$\|\mu_{\eta,j}^{(l)}\|_\infty = O(\sqrt{L}).$$

Next we state Assumptions on the dimension of the B-spline basis $L$, $s_0$, and $s_j^{(l)}$. We allow them to depend on $n$ as long as they satisfy the assumptions.

**Assumption L**
**(1)** $n^{-1}s_j^{(l)2}L^3 \log n \to 0$ uniformly in $l$ $(1 \le l \le L)$ and $j$ $(1 \le j \le p)$.
**(2)** $n^{-1}s_j^{(l)}L^4 \log n \to 0$ uniformly in $l$ $(1 \le l \le L)$ and $j$ $(1 \le j \le p)$.
**(3)** $n^{-1}s_0^2L^4(\log n)^2 \to 0$.

**Lemma 3** *Suppose that Assumptions VC, S2, and E hold and that $n^{-1}L^2 \log n \to 0$. Then for some large constant C, we have*

$$\mathrm{P}_\infty(n^{-1}\boldsymbol{W}_{-j}^T\eta_j^{(l)}) < C\sqrt{\frac{L^2 \log n}{n}}$$

*uniformly in $l$ $(1 \le l \le L)$ and $j$ $(1 \le j \le p)$ with probability tending to 1.*

The convergence rate is worse than that in Lemma 1. This is due to the structure of $\boldsymbol{W}$, (28), and Assumption E. There may be possibility of improvement in this convergence rate. See Remark 4 at the end of this section.

**Lemma 4** *Define* $\widehat{\Sigma}_{-j,-j}$ *as* $\widehat{\Sigma}_{-j,-j} := \frac{1}{n}\boldsymbol{W}_{-j}^T\boldsymbol{W}_{-j}$. *Then suppose that Assumptions VC, S1, and L(1) hold. Then*

$$2\phi^2_{\widehat{\Sigma}_{-j,-j}}(\mathcal{S}_j^{(l)}, 3) \geq \phi^2_{\Sigma_{-j,-j}}(\mathcal{S}_j^{(l)}, 3) \geq \lambda_{\min}(\Sigma)$$

*uniformly in* $l$ $(1 \leq l \leq L)$ *and* $j$ $(1 \leq j \leq p)$ *with probability tending to 1.*

**Proposition 2** *Suppose that Assumptions VC, S1, S2, E, and L(1) hold and take* $\lambda_j^{(l)} = C\sqrt{n^{-1}L^2 \log n}$ *for sufficiently large* $C$. *Then we have*

$$\frac{1}{n}\|\boldsymbol{W}_{-j}(\widehat{\gamma}_j^{(l)} - \gamma_j^{(l)})\|^2 \leq 18\frac{\lambda_j^{(l)2} s_j^{(l)}}{\lambda_{\min}(\Sigma)} \quad \text{and} \quad \mathrm{P}_1(\widehat{\gamma}_j^{(l)} - \gamma_j^{(l)}) \leq 24\frac{\lambda_j^{(l)} s_j^{(l)}}{\lambda_{\min}(\Sigma)}$$

*uniformly in* $l$ $(1 \leq l \leq L)$ *and* $j$ $(1 \leq j \leq p)$ *with probability tending to 1.*

Actually $C$ in Proposition 2 can depend on $l$ and $j$ if it belongs to some suitable interval. Note that $C$ in the definition of $\lambda_j^{(l)}$ is from Lemma 3.

• **Results on** $\widehat{\boldsymbol{b}}$ : We present Propositions 3-5. Hereafter we assume the conditions on the tuning parameters $\lambda_0$ and $\lambda_j^{(l)}$ in Propositions 1 and 2.

**Proposition 3** *Suppose that Assumptions VC, G , S1, S2, E, and L(1)-(3) hold. Then we have*

$$\|\Delta_{1,j}\| < C\frac{1}{n^{1/2}} \cdot \frac{s_0 L^2 \log n}{n^{1/2}}$$

*uniformly in* $j$ $(1 \leq j \leq p)$ *with probability tending to 1 for sufficiently large* $C$.

**Proposition 4** *Suppose that Assumptions VC, G , S1, S2, E, and L(1)(2) hold. Then we have*

$$\|\Delta_{2,j}\| < C \cdot L^{-3}\Big(\sum_{l=1}^{L} s_j^{(l)}\Big)^{1/2} \log n \leq CL^{-5/2} \log n (\max_{l,j} s_j^{(l)})^{1/2}$$

*uniformly in* $j$ $(1 \leq j \leq p)$ *with probability tending to 1 for sufficiently large* $C$.

We give a comment on possibility of some improvements on Proposition 4 in Remark 5 at the end of this section. It is just a conjecture that we have not proved yet.

We introduce some more notation before Proposition 5. We denote the residual vectors from the group Lasso in (13) by $\widehat{\eta}_j^{(l)} := W_j - \boldsymbol{W}_{-j}^T\widehat{\gamma}_j^{(l)} \in \mathbb{R}^n$ and note that $\widehat{E}_j = (\widehat{\eta}_j^{(1)}, \ldots, \widehat{\eta}_j^{(L)})$, where $\widehat{E}_j$ is an $n \times L$ matrix. Besides, we set

$$\begin{aligned}\widehat{\Omega} := \widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^T &= \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T\boldsymbol{W}\widehat{\Theta}^T \qquad (29)\\ &= \{\mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2})\}^T\frac{1}{n}(\widehat{E}_1, \ldots, \widehat{E}_p)^T(\widehat{E}_1, \ldots, \widehat{E}_p)\mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2})\end{aligned}$$

and define its submatrix $\widehat{\Omega}_{j,k}$ in the same way as $\Sigma_{j,k}$ and $\Theta_{j,k}$ are defined as submatrices of $\Sigma$ and $\Theta$, respectively. We used (16) and (18) in the last line. Note that $\widehat{\Omega}$ is a $(pL) \times (pL)$ matrix and it is the conditional variance matrix of $n^{-1/2}\widehat{\Theta}\boldsymbol{W}^T\epsilon$. Recall $\mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2})$ is the second matrix on the RHS of (18).

**Proposition 5** *Suppose that Assumptions VC, G , S1, S2, E, and L(1)(2) hold and fix a positive integer m. For any* $\{j_1, \dots, j_m\} \subset \{1, \dots, p\}$*, we define a symmetric matrix* $\Delta$ *as*

$$\Delta := \begin{pmatrix} \widehat{\Omega}_{j_1,j_1} & \cdots & \widehat{\Omega}_{j_1,j_m} \\ \vdots & \ddots & \vdots \\ \widehat{\Omega}_{j_m,j_1} & \cdots & \widehat{\Omega}_{j_m,j_m} \end{pmatrix} - \begin{pmatrix} \Theta_{j_1,j_1} & \cdots & \Theta_{j_1,j_m} \\ \vdots & \ddots & \vdots \\ \Theta_{j_m,j_1} & \cdots & \Theta_{j_m,j_m} \end{pmatrix}.$$

*Then we have*

$$|\lambda_{\min}(\Delta)| \vee |\lambda_{\max}(\Delta)| \to 0$$

*uniformly in* $\{j_1, \dots, j_m\}$ *with probability tending to 1.*

Our main result, Theorem 1, immediately follows from Propositions 3-5. Recall that $\Delta_1 = (\Delta_{1,1}^T, \dots, \Delta_{1,p}^T)^T$ and $\Delta_2 = (\Delta_{2,1}^T, \dots, \Delta_{2,p}^T)^T$. We give a comment on spline bases in Remark 2 below.

**Theorem 1** *Suppose that Assumptions VC, G , S1, S2, E, and L(1)-(3) hold. Then the debiased estimator is represented as*

$$\widehat{\boldsymbol{b}} - \boldsymbol{\beta}_0 = \frac{1}{n}\widehat{\Theta}\boldsymbol{W}\boldsymbol{\epsilon} - \Delta_1 + \Delta_2$$

*and we have*

$$\|\Delta_{1,j}\| = o_p(n^{-1/2}) \quad \text{and} \quad \|\Delta_{2,j}\| < C \log n \cdot L^{-5/2}(\max_{l,j} s_j^{(l)})^{1/2}$$

*uniformly in* $j$ $(1 \le j \le p)$ *with probability tending to 1 for sufficiently large C. Besides, we have* $n^{-1/2}\widehat{\Theta}\boldsymbol{W}^T\boldsymbol{\epsilon} \,|\, \{\underline{X}_i, Z_i\}_{i=1}^n \sim \mathrm{N}(0, \sigma_\epsilon^2\widehat{\Omega})$ *and* $\widehat{\Omega}$ *converges in probability to* $\Theta$ *blockwise as defined in Proposition 5.*

**Remark 2** This remark concerns other spline bases. We can take another spline basis $B'(z)$ satisfying $B'(z) = AB(z)$ and $C_1 < \lambda_{\min}(AA^T) \le \lambda_{\max}(AA^T) < C_2$ for some positive constants $C_1$ and $C_2$. For example, an orthonormal basis $B'(z)$ satisfying $\int B'(z)(B'(z))^T dz = I_L$. This is because we deal with and evaluate everything blockwise. We use the desirable properties of the B-spline basis in the proofs and then we should apply the conformable linear transformation blockwise.

We consider applications of Theorem 1. Recall we have $\max_{j \in \mathcal{S}_0} \|g_j - B^T\beta_{0j}\|_\infty = O(L^{-3})$ by Assumption G.

**• Statistical inference under the original high-dimensional model**

**(1)** $\|g_j\|_2$ : Suppose we use a spline basis satisfying the orthonormal property in Remark 2. Then $\|\widehat{b}_j\|$ is the estimator of $\|g_j\|_2$. We can also deal with $\|g_j - g_k\|_2$ and then $\|\widehat{b}_j - \widehat{b}_k\|$

is the estimator. Recall again that the SCAD gives no information of $\|g_j\|_2$ when this $j$ is not selected. Most of screening procedures rely on an assumption like the one that marginal models faithfully reflect the true model. It is important to have a de-biased estimator of $\|g_j\|_2$ for any $j$ based on the initial and original high-dimensional varying coefficient model (1).

Theorem 1 suggests that for any fixed $j$,

$$\|\widehat{b}_j - \beta_{0j}\| = O_p\Big(\sqrt{\frac{L}{n}}\Big)$$

if $\sqrt{n^{-1}L}/\{\log n \cdot L^{-5/2}(\max_{l,j} s_j^{(l)})^{1/2}\} \to \infty$. This reduces to $L^6/\{n(\log n)^2 \max_{l,j} s_j^{(l)}\} \to \infty$. Note that $\|\beta_{0j}\| - \|g_j\|_2 = O(L^{-3})$ uniformly in $j$ under Assumption G and this approximation error is negligible compared to $(L/n)^{1/2}$.

In addition to point estimation of $\|g_j\|_2$, we can carry out hypothesis testing of $H_0 : \|g_j\|_2 = 0$ vs. $H_1 : \|g_j\|_2 \neq 0$ for any $j$. Then we can approximate the distribution of $\|\widehat{b}_j\|$ by bootstrap for $j \notin \mathcal{S}_0$ to compute the critical value as we did in the simulation studies.

**(2)** $g_j(z)$ : We estimate $g_j(z)$ with $B^T(z)\widehat{b}_j$. Since $B^T(z)\beta_{0j} - g_j(z) = O(L^{-3})$ under Assumption G, this approximation error is negligible compared to the effect of $\Delta_2$ and $(L/n)^{1/2}$. Note that $\{n^{-1}B^T(z)\widehat{\Omega}_{j,j}B(z)\}^{1/2} \sim (L/n)^{1/2}$ in probability. Therefore for any fixed $j$, we have

$$n^{1/2}B^T(z)(\widehat{b}_j - \beta_{0j})/\{B^T(z)\widehat{\Omega}_{j,j}B(z)\}^{1/2} \xrightarrow{d} \mathrm{N}(0, \sigma_\epsilon^2) \tag{30}$$

if $\sqrt{n^{-1}L}/\{\log n \cdot L^{-2}(\max_{l,j} s_j^{(l)})^{1/2}\} \to \infty$. This reduces to $L^5/\{n(\log n)^2 \max_{l,j} s_j^{(l)}\} \to \infty$. This condition may be a little restrictive. However, a smaller $L$ may work practically from Remarks 4 and 5 below. See Subsection S.2.3 in the Supplement for some numerical examples of confidence bands for $g_j(z)$.

We state some remarks here. Those remarks are about possible improvements of our assumptions and we have not proved them yet.

**Remark 3** This remark is about Assumption E. First we consider the case of $l = 1$ for simplicity of notation. For $l = 1$, $\mu_{\eta,j}^{(1)}(Z_i)$ and $\sigma_{\eta,j}^{(1)2}(Z_i)$, $i = 1, \ldots, n$, are written as

$$\mu_{\eta,j}^{(1)}(Z_i) = a_j^{(1)T}\{\mu_X(Z_i) \otimes B(Z_i)\} \quad \text{and} \quad \sigma_{\eta,j}^{(1)2}(Z_i) = a_j^{(1)T}[\Sigma_X(Z_i) \otimes \{B(Z_i)\}^{\otimes 2}]a_j^{(1)},$$

where $a_j^{(1)} := (1, 0, \ldots, 0, -\gamma_j^{(1)T})^T \in \mathbb{R}^{pL}$ and $\|a_j^{(1)}\| = O(1)$ uniformly in $j$ from Assumption S2. (28) easily follows from the local support property of $B(z)$. This holds for the other $l$. On the other hand, $\mu_{\eta,j}^{(l)}(Z_i)$ is rewritten for general $l$ as

$$\mu_{\eta,j}^{(l)}(Z_i) = \mu_{X,j}(Z_i)B_l(Z_i) - \sum_{s \in \mathcal{S}_j^{(l)}} \mu_{X,s}(Z_i)b_{s,j}^{(l)T}B(Z_i) \quad \text{and} \quad \sum_{s \in \mathcal{S}_j^{(l)}} \|b_{s,j}^{(l)}\|^2 = \|\gamma_j^{(l)}\|^2,$$

where $b^{(l)}_{s,j}$ is part of $\gamma^{(l)}_j$. If $\sum_{s\in\mathcal{S}^{(l)}_j}\|b^{(l)}_{s,j}\| < C$ or $s^{(l)}_j < C$ uniformly in $l$ and $j$ for some positive constant $C$, Assumption E holds because of the local support property of the B-spline basis. Besides since only a finite number of elements of $B(z)$ are not zero for any $z$ due to its local support property, Assumption E seems to be a reasonable one.

**Remark 4** This remark refers to possible improvement on Lemma 3. In Lemma 3, we should evaluate the expression inside the expectation on the LHS of (31).

$$\begin{aligned}\mathrm{E}\Big[\sum_{m=1}^{L}\Big\{\frac{1}{n}\sum_{i=1}^{n}X_{i,k}B_m(Z_i)\eta^{(l)}_{i,j}\Big\}^2\Big] &= \frac{1}{n}\mathrm{E}\Big\{(X_{1,k}\eta^{(l)}_{1,j})^2\sum_{m=1}^{L}B^2_m(Z_1)\Big\} \le C_1\frac{L}{n}\mathrm{E}\{(X_{1,k}\eta^{(l)}_{1,j})^2\} \\ &\le \frac{C_1L}{n}\mathrm{E}\{|X_{1,k}|^{2p_1}\}^{1/p_1}\mathrm{E}\{|\eta^{(l)}_{1,j}|^{2p_2}\}^{1/p_2}\end{aligned} \tag{31}$$

for some positive constant $C_1$ and $(p_1, p_2)$ satisfying $1/p_1 + 1/p_2 = 1$. Note that we used Assumption VC and the fact for some positive constant $C_3$, $\sum_{m=1}^{L} B^2_m(Z_1) < C_3L$ uniformly in $Z_1$ here. If we take $p_1 = 4$ and $p_2 = 4/3$, we have (31)= $O(L^{3/2}/n)$ and this suggests there may be possibility of improvement in convergence rate up to $\sqrt{n^{-1}L^{3/2}\log n}$. We have not proved this conjecture yet.

**Remark 5** This remark is about possible improvement on Proposition 4. Note that

$$\begin{aligned}\begin{pmatrix}\Delta_{2,1}\\ \vdots \\ \Delta_{2,p}\end{pmatrix}^T &= \frac{1}{n}r^T\boldsymbol{W}\begin{pmatrix} I_L & -\widehat{\Gamma}_{2,1} & \cdots & -\widehat{\Gamma}_{p,1} \\ -\widehat{\Gamma}_{1,2} & I_L & \cdots & -\widehat{\Gamma}_{p,2} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\Gamma}_{1,p} & -\widehat{\Gamma}_{2,p} & \cdots & I_L\end{pmatrix}\mathrm{diag}(T_1^{-2},\ldots,T_p^{-2}) \\ &= \frac{1}{n}r^T(\widehat{E}_1,\ldots,\widehat{E}_p)\mathrm{diag}(T_1^{-2},\ldots,T_p^{-2})\end{aligned}$$

and

$$\frac{1}{n}r^T\widehat{\eta}^{(l)}_j = \frac{1}{n}r^T\eta^{(l)}_j + \frac{1}{n}r^T\boldsymbol{W}_{-j}(\widehat{\gamma}^{(l)}_j - \gamma^{(l)}_j).$$

Recall the definition of $\widehat{E}_j$ in (15) and $\widehat{E}_j = (\widehat{\eta}^{(1)}_j, \ldots, \widehat{\eta}^{(L)}_j)$. Since

$$\begin{aligned}|n^{-1}r^T\boldsymbol{W}_{-j}(\widehat{\gamma}^{(l)}_j - \gamma^{(l)}_j)| &\le (n^{-1}\|r\|^2)^{1/2}(n^{-1}\|\boldsymbol{W}_{-j}(\widehat{\gamma}^{(l)}_j - \gamma^{(l)}_j)\|^2)^{1/2} \\ &\le CL^{-3}(\log n)^{1/2}(\max_{l,j} s^{(l)}_j)^{1/2}\sqrt{\frac{L^2\log n}{n}}\end{aligned}$$

uniformly in $j$ with probability tending to 1 for some positive constant $C$, this is small enough. Hence we have only to evaluate $n^{-1}r^T\eta^{(l)}_j$. Recalling $r_i = \sum_{j\in\mathcal{S}_0} X_{i,j}(g_j(Z_i) - B^T(Z_i)\beta_{0j})$ and $\eta^{(l)}_j = W^{(l)}_j - \boldsymbol{W}_{-j}\gamma^{(l)}_j$, we conjecture that $n^{-1}r^T\eta^{(l)}_j$ is much smaller than $O_p(L^{-3})$ given in the proof of the proposition. We have not proved this conjecture yet.

# 4 Numerical studies

In this section, we present the results of simulation studies. The proposed de-biased group Lasso estimator may look complicated. However, it worked well in the simulation studies and the results imply that this de-biased group Lasso estimator is quite promising.

In the studies, we present the results on hypothesis testing of whether $\|g_j\|_2 = 0$ or not for $j = 1, \dots, 12$ in Models 1-3 defined below. We also present some more simulation results and a real data application in Section S.2 in the Supplement.

We used the cv.gglasso function of the R package 'gglasso' version 1.4 on R x64 3.5.0. The package is provided by Profs Yi Yang and Hui Zou. See [32] for more details. We chose tuning parameters by using the CV procedure of the cv.gglasso function. First we computed $\widehat{\boldsymbol{\beta}}$ by using the CV procedure and then corrected the bias of it to get $\widehat{\boldsymbol{b}}$. We also used the CV procedure when we computed $\widehat{\boldsymbol{\Theta}}$. We didn't optimize $\widehat{\boldsymbol{b}}$ with respect to $\lambda_0$ because it took too much of time even for one repetition. We used an orthonormal spline basis which is constructed from the quadratic equispaced B-spline basis.

In the three models, $Z_i$ follows the uniform distribution on $[0,1]$ $X_{i,1} \equiv 1$, and $\{X_{i,j}\}_{j=2}^p$ follows a stationary Gaussian AR(1) process with $\rho = 0.5$. We took $\mathrm{E}\{X_{i,2}\} = 0$ and $\mathrm{E}\{X_{i,2}^2\} = 1$ and $Z_i$ and $\{X_{i,j}\}_{j=2}^p$ are mutually independent. As for the error term, we took $\epsilon_i \sim \mathrm{N}(0,3)$. We tried two cases, $(L, p, n, \text{Repetition number}) = (5, 250, 250, 200)$ and $(5, 350, 350, 200)$. Note that the actual dimension is $pL = 1250$ and $1750$. Besides, the tuning parameters were determined by the data and one iteration needs 61 runs of the group Lasso with very many covariates. Therefore it took a long time for only one case of each model.

In Model 1, we set

$$g_2(z) = 2 + 2\sin(\pi z),\ g_4(z) = 2(2z-1)^2 - 2,\ g_6(z) = 1.8\log(z + 1.718282),\ g_8(z) = 2.5(1-z).$$

All the other functions are set to be 0 and irrelevant.

In Model 2, we set

$$g_2(z) = 2 + 2(2z-1)^3, \quad g_4(z) = 2\cos(\pi z), \quad g_6(z) = \frac{1.8}{1+z^2}, \quad g_8(z) = \frac{\exp(1+z)}{3.4}.$$

All the other functions are set to be 0 and irrelevant.

In Model 3, we set

$$\begin{aligned} &g_2(z) = 2 + 2\sin(\pi z), \quad g_4(z) = 2(2z-1)^2 - 2, \quad g_6(z) = \frac{1.8}{1+z^2}, \\ &g_8(z) = \frac{\exp(1+z)}{3.4}, \quad g_{10}(z) = 1.8\log(z + 1.718282), \quad g_{12}(z) = 2\cos(\pi z). \end{aligned}$$

All the other functions are set to be 0 and irrelevant.

We considered hypothesis testing of

$$H_0 : ||g_j||_2 = 0 \quad \text{vs.} \quad H_1 : ||g_j||_2 > 0 \tag{32}$$

for $j = 1, \dots, 12$ in Models 1-3. We computed the critical values from the result that $\sqrt{n}(\widehat{b}_j - \beta_{0j})$ is approximately distributed as $\mathrm{N}(0, \widehat{\Omega}_{j,j})$ in Theorem 1. Then $||\widehat{b}_j||^2$ is the estimator of $||g_j||_2$ since we used an orthonormal B-spline basis here. We compared $||\widehat{b}_j||^2$ and the simulated critical value. The nominal significance levels are 0.05 and 0.10.

In Tables 1-12, each entry is the rate of rejecting $H_0$. Tables 1, 3, 5, 7, 9, and 11 are for relevant $j$ ($H_1$ is true) and Tables 2, 4, 6, 8, 10, and 12 are for irrelevant $j$ ($H_0$ is true).

As shown in Tables for relevant covariates ($H_1$), the rejection rate is 1.00 for any case. As for irrelevant covariates ($H_0$), the actual significance levels are close to the nominal ones except for $j = 7$ in Models 1 and 2 and $j = 7, 9$ in Model 3. Note that the standard errors are 0.022($\alpha = 0.10$) and 0.016($\alpha = 0.05$) since the repetition number is 200 due to the long computational time. We also tried 6 more cases where every $g_j(z)$ is replaced with $g_j(z)/\sqrt{2}$. There is no significant differences and the results of the 6 cases are presented in the supplement. These simulation results imply that our de-biased Lasso procedure is very promising for statistical inference under the original high-dimensional model, i.e. statistical inference without variable selection.

Table 1: $H_1$ for Model 1 with $p = 250$ and $n = 250$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table 2: $H_0$ for Model 1 with $p = 250$ and $n = 250$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.10 | 0.06 | 0.06 | 0.18 | 0.12 | 0.08 | 0.15 | 0.08 |
| $\alpha = 0.05$ | 0.06 | 0.02 | 0.02 | 0.13 | 0.06 | 0.04 | 0.08 | 0.06 |

Table 3: $H_1$ for Model 2 with $p = 250$ and $n = 250$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table 4: $H_0$ for Model 2 with $p = 250$ and $n = 250$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.11 | 0.11 | 0.18 | 0.18 | 0.12 | 0.08 | 0.14 | 0.12 |
| $\alpha = 0.05$ | 0.06 | 0.06 | 0.10 | 0.11 | 0.06 | 0.04 | 0.08 | 0.05 |

Table 5: $H_1$ for Model 3 with $p = 250$ and $n = 250$

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 6: $H_0$ for Model 3 with $p = 250$ and $n = 250$

| $j$ | 1 | 3 | 5 | 7 | 9 | 11 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.12 | 0.07 | 0.05 | 0.22 | 0.22 | 0.15 |
| $\alpha = 0.05$ | 0.07 | 0.04 | 0.03 | 0.14 | 0.16 | 0.10 |

Table 7: $H_1$ for Model 1 with $p = 350$ and $n = 350$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table 8: $H_0$ for Model 1 with $p = 350$ and $n = 350$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.10 | 0.03 | 0.05 | 0.16 | 0.11 | 0.07 | 0.10 | 0.08 |
| $\alpha = 0.05$ | 0.06 | 0.02 | 0.02 | 0.12 | 0.06 | 0.05 | 0.06 | 0.05 |

Table 9: $H_1$ for Model 2 with $p = 350$ and $n = 350$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table 10: $H_0$ for Model 2 with $p = 350$ and $n = 350$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.09 | 0.10 | 0.10 | 0.16 | 0.11 | 0.06 | 0.11 | 0.08 |
| $\alpha = 0.05$ | 0.04 | 0.04 | 0.06 | 0.10 | 0.06 | 0.04 | 0.06 | 0.05 |

Table 11: $H_1$ for Model 3 with $p = 350$ and $n = 350$

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 12: $H_0$ for Model 3 with $p = 350$ and $n = 350$

| $j$ | 1 | 3 | 5 | 7 | 9 | 11 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.09 | 0.05 | 0.07 | 0.22 | 0.20 | 0.10 |
| $\alpha = 0.05$ | 0.07 | 0.03 | 0.02 | 0.17 | 0.14 | 0.07 |

# 5 Proofs of theoretical results

In this section, we prove Propositions 3-5. We state two technical lemmas before we prove the propositions. These lemmas will be verified in the Supplement.

We define $L \times L$ matrices $\widehat{B}_{j,k}$ and $B_{j,k}$ for $j = 1, \dots, p$ and $k = 1, \dots, p$ as

$$\widehat{B}_{j,k} := \frac{1}{n} \widehat{E}_j^T \widehat{E}_k \quad \text{and} \quad B_{j,k} := \frac{1}{n} \mathrm{E}(E_j^T E_k)$$

See (15) and (9) for the definitions of $\widehat{E}_j$ and $E_j$. Note that and

$$B_{j,j} = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j} = \Theta_{j,j}^{-1} \quad \text{and} \quad B_{j,k} = \mathrm{E}\left\{ \begin{pmatrix} \eta_{1,j}^{(1)} \\ \vdots \\ \eta_{1,j}^{(L)} \end{pmatrix} (\eta_{1,k}^{(1)}, \dots, \eta_{1,k}^{(L)}) \right\}. \tag{33}$$

We establish the convergence of $\widehat{B}_{j,k}$ to $B_{j,k}$ in Lemma 5.

**Lemma 5** *Suppose that Assumptions VC, S1, S2, E, and L(1)(2) hold. Then $\|\widehat{B}_{j,k} - B_{j,k}\|_F \to 0$ uniformly in $j\,(1 \le j \le p)$ and $k\,(1 \le k \le p)$ with probability tending to 1.*

In the next lemma, we establish the desirable properties of $T_j^2$. Recall that $\rho(A)$ is the spectral norm of a matrix $A$.

**Lemma 6** *Suppose that Assumptions VC, S1, S2, E, and L(1)(2) hold. Then we have (a) and (b).*
*(a) For some positive constants $C_1$ and $C_2$, we have $C_1 < \rho(T_j^2) = \rho(T_j^{2T}) < C_2$ and $1/C_2 < \rho(T_j^{-2}) = \rho(T_j^{-2T}) < 1/C_1$ uniformly in $j\,(1 \le j \le p)$ with probability tending to 1.*
*(b) $\|T_j^2 - \Theta_{j,j}^{-1}\|_F \to 0$ and $\sup_{\|x\|=1} \|(T_j^{-2} - \Theta_{j,j})x\| \to 0$ uniformly in $j\,(1 \le j \le p)$ with probability tending to 1.*

Now we begin to prove Propositions 3-5

**Proof of Proposition 3)** Since (21) and the properties of $\kappa_{j,k}^{(l)}$ below (14) imply

$$\Delta_{1,j} = T_j^{-2T} \Lambda_j \sum_{k \neq j} K_{j,k}^T (\widehat{\beta}_k - \beta_{0k})$$

and

$$|\lambda_j^{(l)} \sum_{k \neq j} \kappa_{j,k}^{(l)T} (\widehat{\beta}_k - \beta_{0k})| \le \max_{a,b} \lambda_a^{(b)} \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

we have uniformly in $j$,

$$\|\Delta_{1,j}\| \le \max_{a,b} \lambda_a^{(b)} \rho(T_j^{-2}) L^{1/2} \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{34}$$

Recall that $\max_{a,b} \lambda_a^{(b)} = O(\sqrt{n^{-1} L^2 \log n})$ in Proposition 2. By (34), Lemma 6, and the bound of $\mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ from Proposition 1, we have

$$\|\Delta_{1,j}\| \le C \lambda_0 s_0 \sqrt{\frac{L^3 \log n}{n}} \tag{35}$$

uniformly in $j$ with probability tending to 1 for some positive constant $C$.

The desired result follows from (35) and the condition on $\lambda_0$ in Proposition 1. Hence the proof of the proposition is complete.

**Proof of Proposition 4)** Write

$$\begin{aligned}
(\Delta_{2,1}^T, \ldots, \Delta_{2,p}^T) &= (n^{-1} \boldsymbol{W}^T r)^T \widehat{\Theta}^T \\
&= (n^{-1} \boldsymbol{W}^T r)^T \begin{pmatrix} I_L & -\widehat{\Gamma}_{2,1} & -\widehat{\Gamma}_{3,1} & \cdots & -\widehat{\Gamma}_{p,1} \\ -\widehat{\Gamma}_{1,2} & I_L & -\widehat{\Gamma}_{3,2} & \cdots & -\widehat{\Gamma}_{p,2} \\ -\widehat{\Gamma}_{1,3} & -\widehat{\Gamma}_{2,3} & I_L & \cdots & -\widehat{\Gamma}_{p,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\widehat{\Gamma}_{1,p} & -\widehat{\Gamma}_{2,p} & -\widehat{\Gamma}_{3,p} & \cdots & I_L \end{pmatrix} \mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2}).
\end{aligned}$$

The above expression implies that the absolute value of the $l$ th element of $T_j^{2T}\Delta_{2,j}$ is bounded from above by

$$\mathrm{P}_\infty(n^{-1}\boldsymbol{W}^T r)(\mathrm{P}_1(\widehat{\gamma}_j^{(l)})+1) \le C_1\mathrm{P}_\infty(n^{-1}\boldsymbol{W}^T r)(s_j^{(l)})^{1/2}(\|\gamma_j^{(l)}\|+1) \tag{36}$$

uniformly in $l$ and $j$ with probability tending to 1 for some positive constant $C_1$. Note that the difference between $\mathrm{P}_1(\widehat{\gamma}_j^{(l)})$ and $\mathrm{P}_1(\gamma_j^{(l)})$ is negligible by Proposition 2 and that $\mathrm{P}_1(\gamma_j^{(l)}) \le (s_j^{(l)})^{1/2}\|\gamma_j^{(l)}\|$.

Thus by Assumption S2(1), (36) and Lemma 6, we have

$$\|\Delta_{2,j}\| \le C_2\mathrm{P}_\infty(n^{-1}\boldsymbol{W}^T r)\Big(\sum_{l=1}^{L} s_j^{(l)}\Big)^{1/2} \tag{37}$$

uniformly in $j$ with probability tending to 1 for some positive constant $C_2$.

(37) and Lemma 1 yield the desired result. Hence the proof of the proposition is complete.

**Proof of Proposition 5)** The desired result follows from (a) and (b) below, which will be verified later in the proof.

(a) For any $x \in \mathbb{R}^L$ and $y \in \mathbb{R}^L$ satisfying $\|x\| = 1$ and $\|y\| = 1$,

$$|x^T(\widehat{\Omega}_{j,k} - \Theta_{j,j}B_{j,k}\Theta_{k,k})y| \to 0$$

uniformly in $x$, $y$, $j$, and $k$ with probability tending to 1.

(b) $\Theta_{j,j}B_{j,k}\Theta_{k,k} = \Theta_{j,k}$

Actually (a) and (b) imply that for any $x \in \mathbb{R}^{mL}$ satisfying $\|x\| = 1$, $x^T\Delta x \to 0$ uniformly in $x$ and $\{j_1, \dots, j_m\}$ with probability tending to 1.

Now we demonstrate (a) and (b).

(a) Recall that $\widehat{\Omega}_{j,k} = T_j^{-2T}\widehat{B}_{j,k}T_k^{-2}$ in (29) and $B_{j,j} = \Theta_{j,j}^{-1}$. Then note that

$$\begin{aligned} x^T(\widehat{\Omega}_{j,k} - \Theta_{j,j}B_{j,k}\Theta_{k,k})y &= \{x^T(T_j^{-2} - \Theta_{j,j})^T\}\widehat{B}_{j,k}T_k^{-2}y \\ &\quad + x^T\Theta_{j,j}(\widehat{B}_{j,k} - B_{j,k})T_k^{-2}y + x^T\Theta_{j,j}B_{j,k}\{(T_k^{-2} - \Theta_{k,k})y\}. \end{aligned} \tag{38}$$

By Lemmas 5 and 6, we have with probability tending to 1,

$$\|\widehat{B}_{j,k} - B_{j,k}\|_F \to 0 \text{ uniformly in } j \text{ and } k \tag{39}$$

and

$$\|(T_j^{-2} - \Theta_{j,j})x\| \to 0 \text{ uniformly in } j \text{ and } x, \tag{40}$$

where $x \in \mathbb{R}^L$ and $\|x\| = 1$.

Besides, by Lemmas 5 and 6, Assumptions S1 and S2(see (25)), (33), and the Cauchy-Schwarz inequality, we have

$$\rho(T_j^{-2}) \le C_1 \tag{41}$$

$$|x^T B_{j,k} y| \le (x^T \Theta_{j,j}^{-1} x)^{1/2} (y^T \Theta_{k,k}^{-1} y)^{1/2} \le C_2 \|x\| \|y\| \tag{42}$$

$$|x^T \widehat{B}_{j,k} y| \le (x^T \widehat{B}_{j,j} x)^{1/2} (y^T \widehat{B}_{k,k} y)^{1/2} \le C_3 \|x\| \|y\| \tag{43}$$

uniformly $j$ and $k$ with probability tending to 1 for some positive constants $C_1$, $C_2$, and $C_3$.

We can evaluate the first term on the RHS of (38) uniformly by using (40), (41), and (43). We can treat the other two terms on the RHS of (38) similarly. We use (S.1) in the Supplement for the second term to show that the absolute value is less than or equal to $\|\Theta_{j,j}^T x\| \|\widehat{B}_{j,k} - B_{j,k}\|_F \|T_k^{-2} y\|$. Hence we have established (a).

(b) When $j = k$, the equation is trivial. First we consider the case of $p = 2$. Take two $L$-dimensional random vectors $U_1$ and $U_2$ satisfying

$$\mathrm{E}\left\{ \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} (U_1^T U_2^T) \right\} = \Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}.$$

We have

$$\Sigma^{-1} = \begin{pmatrix} \Theta_{1,1} & \Theta_{1,2} \\ \Theta_{2,1} & \Theta_{2,2} \end{pmatrix} = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}^{-1}.$$

and consider $U_1 - \Gamma_1^T U_2$ and $U_2 - \Gamma_2^T U_1$, where $\Gamma_1^T = \Sigma_{1,2}\Sigma_{2,2}^{-1}$ and $\Gamma_2^T = \Sigma_{2,1}\Sigma_{1,1}^{-1}$. Then we have

$$\Theta_{1,1} = [\mathrm{E}\{(U_1 - \Gamma_1^T U_2)(U_1 - \Gamma_1^T U_2)^T\}]^{-1} \quad \text{and} \quad \Theta_{2,2} = [\mathrm{E}\{(U_2 - \Gamma_2^T U_1)(U_2 - \Gamma_2^T U_1)^T\}]^{-1}.$$

In addition,

$$B_{1,2} = \mathrm{E}\{(U_1 - \Gamma_1^T U_2)(U_2 - \Gamma_2^T U_1)^T\} = -\Sigma_{1,2}\Sigma_{2,2}^{-1}(\Sigma_{2,2} - \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}) = -\Sigma_{1,2}\Sigma_{2,2}^{-1}\Theta_{2,2}^{-1}. \tag{44}$$

Then (44) and (A-74) in [12] yield

$$\Theta_{1,1} B_{1,2} \Theta_{2,2} = -\Theta_{1,1}\Sigma_{1,2}\Sigma_{2,2}^{-1} = \Theta_{1,2}. \tag{45}$$

Hence we have verified (b) for $p = 2$.

Next we deal with the cases of $p > 2$. We can consider the case of $j = 1$ and $k = 2$ without loss of generality. Take $p$ $L$-dimensional random vectors $U_1, \dots, U_p$ satisfying

$$\mathrm{E}\left\{ \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix} (U_1^T, \dots, U_p^T) \right\} = \Sigma.$$

We define a set of $\Theta_{1,1}, \Theta_{1,2}, \Theta_{2,2}, B_{1,2}$ for this $\Sigma$ by using $U_1, \ldots, U_p$.

Next take the orthogonal projections of $U_1$ and $U_2$ to the linear space spanned by $U_3, \ldots, U_p$ and denote them by $\overline{U}_1$ and $\overline{U}_2$, respectively. We define the residuals $\widehat{U}_1$ and $\widehat{U}_2$ as $\widehat{U}_1 = U_1 - \overline{U}_1$ and $\widehat{U}_2 = U_2 - \overline{U}_2$. Then by (A-74) in [12], we have

$$\begin{pmatrix} \Theta_{1,1} & \Theta_{1,2} \\ \Theta_{2,1} & \Theta_{2,2} \end{pmatrix} = \left[ \mathrm{E} \left\{ \begin{pmatrix} \widehat{U}_1 \\ \widehat{U}_2 \end{pmatrix} (\widehat{U}_1^T \widehat{U}_2^T) \right\} \right]^{-1} . \tag{46}$$

This means that we can define another set of $\Theta_{1,1}, \Theta_{1,2}, \Theta_{2,2}, B_{1,2}$ by using $\widehat{U}_1$ and $\widehat{U}_2$. These two sets of $\Theta_{1,1}, \Theta_{1,2}, \Theta_{2,2}$ are equal to each other. This is because the matrix in (46) is the same submatrix of $\Sigma^{-1}$. As for $B_{1,2}$, the residual of $U_1$ from the orthogonal projection of $U_1$ to $U_2, \ldots, U_p$ is the same as the residual of $\widehat{U}_1$ from the orthogonal projection of $\widehat{U}_1$ to $\widehat{U}_2$. This also holds for $U_2$ and $\widehat{U}_2$. Thus two $B_{1,2}$ are equal to each other.

The result for $p = 2$ implies that

$$\Theta_{1,1} B_{1,2} \Theta_{2,2} = \Theta_{1,2}$$

Hence the proof of (b) is complete.

# References

[1] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[2] P. Breheny. *grpreg: Regularization Paths for Regression Models with Grouped Covariates*, 2019. R package version version 3.2-1.

[3] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods Theory and Applications*. Springer, New York, Dordrecht, Heidelberg, London, 2011.

[4] M. Caner and A. B. Kock. Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *J. Econometrics*, 203:143–168, 2018.

[5] M.-Y. Cheng, T. Honda, J. Li, and H. Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Ann. Statist.*, 42:1819–1849, 2014.

[6] M.-Y. Cheng, T. Honda, and J.-T. Zhang. Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 111:1209–1221, 2016.

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.

[8] J. Fan, Y. Ma, and W. Dai. Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.*, 109:1270–1284, 2014.

[9] J. Fan and R. Song. Sure independence screening in generalized linear models with np-dimensionality. *Ann. Statist.*, 38:3567–3604, 2010.

[10] J. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42:819–849, 2014.

[11] J. Fan and W. Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1:179–195, 2008.

[12] W. H. Greene. *Econometric Analysis 7th ed.* Pearson Education, Harlow, 2012.

[13] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. CRC press, Boca Raton, 2015.

[14] T. Honda and W. K. Härdle. Variable selection in cox regression models with varying coefficients. *J. Statist. Plann. Inference*, 148:67–81, 2014.

[15] T. Honda and R. Yabe. Variable selection and structure identification for varying coefficient cox models. *J. Multivar. Anal.*, 161:103–122, 2017.

[16] J. Z. Huang, C. O. Wu, and L. Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica*, 14:763–788, 2004.

[17] C.-K. Ing and T. L. Lai. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, 22:1473–1513, 2011.

[18] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Machine Learning Research*, 15:2869–2909, 2014.

[19] A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *Ann. Statist.*, 46:2593–2622, 2018.

[20] J. Liu, R. Li, and R. Wu. Feature selection for varying coefficient models with ultrahigh dimensional covariates. *J. Amer. Statist. Assoc.*, 109:266–274, 2014.

[21] J. Liu, W. Zhong, and R. Li. A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58:1–22, 2015.

[22] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39:2164–2204, 2011.

[23] R. Mitra and C.-H. Zhang. The benefit of group sparsity in group inference with debiased scaled group lasso. *Electronic J. Statist.*, 10:1829–1873, 2016.

[24] L. L. Schumaker. *Spline Functions: Basic Theory 3rd ed.* Cambridge University Press, Cambridge, 2007.

[25] B. Stucky and S. van de Geer. Asymptotic confidence regions for high-dimensional structured sparsity. *IEEE Trans. Signal Process.*, 66:2178–2190, 2018.

[26] R. J. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. Ser. B*, 58:267–288, 1996.

[27] S. van de Geer. *Estimation and Testing under Sparsity*. Springer, Switzerland, 2016.

[28] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42:1166–1202, 2014.

[29] A. D. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

[30] H. Wang. Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.*, 104:1512–1524, 2009.

[31] F. Wei, J. Huang, and Li. H. Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica*, 21:1515–1540, 2011.

[32] Y. Yang and H. Zou. *gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm*, 2017. R package version 1.4.

[33] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. Ser. B*, 68:49–67, 2006.

[34] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Royal Statist. Soc. Ser. B*, 76:217–242, 2014.

[35] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101:1418–1429, 2006.

# Supplement

## S.1 Technical proofs

In this supplement, we prove all the lemmas and Propositions 1 and 2.

We often appeal to the standard arguments based on Bernstein's inequality and reproduce the inequality from [29] for reference.

**Lemma 7** *(Bernstein's inequality) Let $Y_1, \dots, Y_n$ be independent random variables such that* $\mathrm{E}(Y_i) = 0$ *and* $\mathrm{E}(|Y_i|^m) \le m!M^{m-2}v_i/2$ *for any positive integer $m \ge 2$ and $i = 1, \dots, n$ for some positive constants $M$ and $v_i$. Then we have*

$$\mathrm{P}(|Y_1 + \cdots + Y_n| > x) \le 2\exp\Big\{ -\frac{x^2}{2(v + Mx)} \Big\}$$

*for $v = \sum_{i=1}^n v_i$.*

We explain here why our Assumption VC in Section 3 allows us to use Bernstein's inequality. Since

$$\mathrm{E}\{\exp(\alpha^T \check{\underline{X}}_i)|Z_i\} \le \exp(\|\alpha\|^2\sigma^2/2) \text{ and } \lambda_{\min}(\Sigma_X(Z_i))\|\alpha\|^2 \le \mathrm{Var}(\alpha^T \check{\underline{X}}_i|Z_i),$$

we have

$$\|\alpha\|^2 \le C\mathrm{Var}(\alpha^T \check{\underline{X}}_i|Z_i)$$

for some positive constant $C$ by Assumptions VC(1)-(2). Recall that $\check{\underline{X}}_i$ is defined just above Assumption VC. Hence we can use $\sigma^2\times$the conditional variance instead of $\|\alpha\|^2\sigma^2$ when we evaluate the moments necessary for Bernstein's inequality. Then we can use assumptions and properties of the conditional variances as well as the conditional means. Note that $\alpha$ can depend on $Z_i$.

Besides, we state some inequalities related to the Frobenius norm here.

For any matrices $A$ and $B$ for which $AB$ is defined, we have

$$\|AB\|_F \le \|A\|_F\|B\|_F. \tag{S.1}$$

This implies that for a $k \times k$ symmetric matrix $A$, we have

$$|\lambda_{\min}(A)| \vee |\lambda_{\max}(A)| \le \|A\|_F \tag{S.2}$$

The first one is well known and a requirement of the matrix norms. (S.2) follows from applying (S.1) to $x^TAx$ with $x \in \mathbb{R}^k$ and $\|x\| = 1$.

**Proof of Lemma 1)** Write

$$\frac{1}{n}\sum_{i=1}^{n} W_{i,j}^{(l)}(\epsilon_i + r_i) = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}B_l(Z_i)\epsilon_i + \frac{1}{n}\sum_{i=1}^{n} X_{i,j}B_l(Z_i)r_i := a_{l,j} + b_{l,j}. \tag{S.3}$$

First we evaluate $a_{l,j}$ and $b_{l,j}$ defined in (S.3) and then consider $(\sum_{l=1}^{L} a_{l,j}^2)^{1/2}$ and $(\sum_{l=1}^{L} b_{l,j}^2)^{1/2}$.
Evaluation of $a_{l,j}$ : By Assumption VC and the local support property of the B-spline basis, we have for some positive constants $C_1$ and $C_2$ that

$$\mathrm{E}\{X_{1,j}B_l(Z_1)\epsilon_1\} = 0 \quad \text{and} \quad \mathrm{E}\{|X_{1,j}B_l(Z_1)\epsilon_1|^m\} \le C_1 m!(C_2 L^{1/2})^{m-2}$$

for any positive integer $m \geq 2$ uniformly in $l$ and $j$. By employing the standard argument based on Bernstein's inequality, we obtain

$$|a_{l,j}| \le C_3\sqrt{\frac{\log n}{n}} \tag{S.4}$$

uniformly in $l$ and $j$ with probability tending to 1 for some positive constant $C_3$.
Evaluation of $b_{l,j}$ : By (24) and the non-negativity of the B-spline basis functions, we have

$$|b_{l,j}| \le C_1\frac{(\log n)^{1/2}}{n}\sum_{i=1}^{n} B_l(Z_i)|r_i| \le C_2\frac{\log n}{nL^3}\sum_{i=1}^{n} B_l(Z_i) \tag{S.5}$$

uniformly in $l$ and $j$ with probability tending to 1 for some positive constants $C_1$ and $C_2$. Since for some positive constants $C_3$ and $C_4$,

$$\mathrm{E}\{B_l(Z_1)\} \le C_3 L^{-1/2} \quad \text{and} \quad \mathrm{E}\{B_l^m(Z_1)\} \le C_4 L^{(m-2)/2}$$

for any positive integer $m \geq 2$ uniformly in $l$, we can apply the standard argument based on Bernstein's inequality and get

$$\frac{1}{n}\sum_{i=1}^{n} B_l(Z_i) \le C_5 L^{-1/2} \tag{S.6}$$

uniformly in $l$ with probability tending to 1 for some positive constant $C_5$. Therefore by (S.5) and (S.6), we have for some positive constant $C_6$,

$$|b_{l,j}| \le C_6 L^{-1/2}\frac{\log n}{L^3} \tag{S.7}$$

uniformly in $l$ and $j$ with probability tending to 1.

(S.4) and (S.7) yield

$$\Big(\sum_{l=1}^{L} a_{l,j}^2\Big)^{1/2} \le C_7\sqrt{\frac{L\log n}{n}} \quad \text{and} \quad \Big(\sum_{l=1}^{L} b_{l,j}^2\Big)^{1/2} \le C_8\frac{\log n}{L^3} \tag{S.8}$$

uniformly in $j$ with probability tending to 1 for some positive constants $C_7$ and $C_8$. Hence the desired results follow from (S.8).

**Proof of Lemma 2)** Set

$$\delta_n := \max_{1 \le s,t \le pL} |(\widehat{\Sigma} - \Sigma)_{s,t}|.$$

Notice that $(\widehat{\Sigma} - \Sigma)_{s,t}$, the $(s,t)$ element of $\widehat{\Sigma} - \Sigma$, is written as

$$\frac{1}{n} \sum_{i=1}^{n} B_{l_1}(Z_i) B_{l_2}(Z_i) X_{i,j_1} X_{i,j_2} - \mathrm{E}\{B_{l_1}(Z_1) B_{l_2}(Z_1) X_{1,j_1} X_{1,j_2}\}.$$

By Assumption VC and the properties of the B-spline basis, we have uniformly in $l_1$, $l_2$, $j_1$, and $j_2$,

$$\mathrm{E}\{|B_{l_1}(Z_1) B_{l_2}(Z_1) X_{1,j_1} X_{1,j_2}|\} \le C_1 \quad \text{and}$$
$$\mathrm{E}\{|B_{l_1}(Z_1) B_{l_2}(Z_1) X_{1,j_1} X_{1,j_2}|^m\} \le \mathrm{E}\{|B_{l_1}(Z_1) X_{1,j_1}|^{2m}\} + \mathrm{E}\{|B_{l_2}(Z_1) X_{1,j_2}|^{2m}\} \le C_2 L (C_3 L)^{m-2} m!$$

for any positive integer $m \ge 2$ for some positive constants $C_1$, $C_2$, and $C_3$. Thus by applying the standard argument based on Bernstein's inequality, we obtain

$$\delta_n \le C_4 \sqrt{\frac{L \log n}{n}} \tag{S.9}$$

with probability tending to 1 for some positive constant $C_4$.

We evaluate $|\boldsymbol{v}^T (\Sigma - \widehat{\Sigma}) \boldsymbol{v}|$ for $\boldsymbol{v} = (v_1^T, \dots, v_p^T)^T \in \Psi(\mathcal{S}_0, 3)$ by employing (S.9). Notice that

$$\| \sum_{k=1}^{p} (\widehat{\Sigma}_{j,k} - \Sigma_{j,k}) v_k \| \le \sum_{k=1}^{p} \|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\|_F \|v_k\| \le \delta_n L \mathrm{P}_1(\boldsymbol{v}).$$

We used (S.1) and (S.9) here. Then

$$\begin{aligned} |\boldsymbol{v}^T (\Sigma - \widehat{\Sigma}) \boldsymbol{v}| &\le \mathrm{P}_1(\boldsymbol{v}) \mathrm{P}_\infty((\Sigma - \widehat{\Sigma}) \boldsymbol{v}) \le \{\mathrm{P}_1(\boldsymbol{v})\}^2 \delta_n L \\ &\le \{\mathrm{P}_1(\boldsymbol{v}_{\mathcal{S}_0}) + \mathrm{P}_1(\boldsymbol{v}_{\overline{\mathcal{S}}_0})\}^2 \delta_n L \le 16 \delta_n L \{\mathrm{P}_1(\boldsymbol{v}_{\mathcal{S}_0})\}^2 \le 16 s_0 \delta_n L \|\boldsymbol{v}_{\mathcal{S}_0}\|^2. \end{aligned}$$

This implies

$$\boldsymbol{v}^T \widehat{\Sigma} \boldsymbol{v} \ge \boldsymbol{v}^T \Sigma \boldsymbol{v} - 16 s_0 \delta_n L \|\boldsymbol{v}_{\mathcal{S}_0}\|^2$$

for $\boldsymbol{v} = (v_1^T, \dots, v_p^T)^T \in \Psi(\mathcal{S}_0, 3)$. Hence

$$\frac{\boldsymbol{v}^T \widehat{\Sigma} \boldsymbol{v}}{\|\boldsymbol{v}_{\mathcal{S}_0}\|^2} \ge \frac{\boldsymbol{v}^T \Sigma \boldsymbol{v}}{\|\boldsymbol{v}_{\mathcal{S}_0}\|^2} - 16 s_0 \delta_n L \ge \frac{\boldsymbol{v}^T \Sigma \boldsymbol{v}}{\|\boldsymbol{v}\|^2} - 16 s_0 \delta_n L. \tag{S.10}$$

The desired result follows from (S.9) and (S.10). Hence the proof of the lemma is complete.

We will prove Proposition 1 a little more generally than stated in Section 3. We assume we have some prior knowledge on $\mathcal{S}_0$, i.e. we know an index set $\mathcal{S}_{prior} \subset \mathcal{S}_0$ and we don't impose any penalties on $\mathcal{S}_{prior}$. This means we replace $\mathrm{P}_1(\boldsymbol{\beta})$ with $\sum_{j\in\overline{\mathcal{S}}_{prior}} \|\beta_j\|$ or $\mathrm{P}_1(\boldsymbol{\beta}_{\overline{\mathcal{S}}_{prior}})$ in (6).

**Proof of Proposition 1)** In the proof, we confine ourselves to this intersection of the two sets :

$$\{\, \mathrm{P}_\infty(n^{-1}\boldsymbol{W}^T(r+\epsilon)) \le \lambda_0/2 \,\} \cap \{\, 2\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3) \ge \phi^2_{\Sigma}(\mathcal{S}_0, 3) \,\}.$$

The former set is related to the deviation condition and the latter one is related to the RE condition. According to Lemma 1 and the condition on $\lambda_0$, the probability of this intersection tends to 1.

Because of the optimality of $\widehat{\boldsymbol{\beta}}$, we have

$$\frac{1}{n}\|Y - \boldsymbol{W}\widehat{\boldsymbol{\beta}}\|^2 + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}} \|\widehat{\beta}_j\| \le \frac{1}{n}\|Y - \boldsymbol{W}\boldsymbol{\beta}_0\|^2 + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}} \|\beta_{0j}\|. \tag{S.11}$$

By (S.11) and the deviation condition, we get

$$\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}} \|\widehat{\beta}_j\| \le \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}\cap\mathcal{S}_0} \|\beta_{0j}\|.$$

Since $\overline{\mathcal{S}}_{prior} = \overline{\mathcal{S}}_0 \cup (\overline{\mathcal{S}}_{prior} \cap \mathcal{S}_0)$, the above inequality reduces to

$$\begin{aligned} &\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_0} \|\widehat{\beta}_j\| \qquad\qquad (\mathrm{S}.12)\\ \le\;& \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}\cap\mathcal{S}_0} \|\widehat{\beta}_j\| + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}\cap\mathcal{S}_0} \|\beta_{0j}\|\\ \le\;& \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + 2\lambda_0 \sum_{j\in\overline{\mathcal{S}}_{prior}\cap\mathcal{S}_0} \|\widehat{\beta}_j - \beta_{0j}\|\\ \le\;& \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + 2\lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}). \end{aligned}$$

This (S.12) is equivalent to

$$\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + 2\lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) \le \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}) + \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) + 2\lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}).$$

The above inequality yields

$$\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) \le 3\lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}) \le 3\lambda_0 s_0^{1/2} \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}\|. \tag{S.13}$$

Note that (S.13) implies that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \in \Psi(\mathcal{S}_0, 3)$ since $\mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) = \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0} - \boldsymbol{\beta}_{0\overline{\mathcal{S}}_0})$. Thus we recall the definition of $\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3)$ and obtain

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + \lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) \\
&\leq \frac{3\lambda_0 s_0^{1/2}}{\phi_{\widehat{\Sigma}}(\mathcal{S}_0, 3)} n^{-1/2}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\| \\
&\leq \frac{1}{2n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + \frac{9\lambda_0^2 s_0}{2\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3)}
\end{aligned}$$

Finally by the RE condition, we have

$$\frac{1}{n}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + 2\lambda_0 \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) \leq \frac{9\lambda_0^2 s_0}{\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3)} \leq \frac{18\lambda_0^2 s_0}{\phi^2_{\Sigma}(\mathcal{S}_0, 3)}. \tag{S.14}$$

The former half of the proposition follows from (S.14).

Next we verify the latter half. Since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \in \Psi(\mathcal{S}_0, 3)$,

$$\mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) \leq 3s_0^{1/2}\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}\|.$$

Thus we have

$$\mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \leq \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}) + 3s_0^{1/2}\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}\| \leq 4s_0^{1/2}\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{0\mathcal{S}_0}\|. \tag{S.15}$$

By (S.15), the definition of $\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3)$, (S.14), and the RE condition, we have

$$\begin{aligned}
&\mathrm{P}_1(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\leq \frac{4s_0^{1/2}}{\phi_{\widehat{\Sigma}}(\mathcal{S}_0, 3)} n^{-1/2}\|\boldsymbol{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\| \leq \frac{12 s_0 \lambda_0}{\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3)} \leq \frac{24 s_0 \lambda_0}{\phi^2_{\Sigma}(\mathcal{S}_0, 3)}.
\end{aligned}$$

This is the latter half of the proposition. Hence the proof of the proposition is complete.

**Proof of Lemma 3)** First we should evaluate

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,k}B_m(Z_i)\eta^{(l)}_{i,j} - \mathrm{E}\{X_{1,k}B_m(Z_1)\eta^{(l)}_{1,j}\}$$

uniformly in $k, m, l, j$. Note that $\mathrm{E}\{X_{1,k}B_m(Z_1)\eta^{(l)}_{1,j}\} = 0$ from the definition of $\eta^{(l)}_{1,j}$. Denote the conditional mean and variance of $X_{1,k}B_m(Z_1)$ given $Z_1$ by $\widetilde{\mu}_{k,m}(Z_1)$ and $\widetilde{\sigma}^2_{k,m}(Z_1)$, respectively and note that $\|\widetilde{\mu}_{k,m}\|_\infty \leq C_1 L^{1/2}$ and $\|\widetilde{\sigma}^2_{k,m}\|_\infty \leq C_2 L$ uniformly in $k$ and $m$ for some positive constants $C_1$ and $C_2$ by Assumption VC. Besides, $\mathrm{E}\{\widetilde{\mu}^2_{k,m}(Z_1) + \widetilde{\sigma}^2_{k,m}(Z_1)\}$ is also uniformly bounded. By Assumptions VC and E, (27), and (28), and some calculations, we have

$$\mathrm{E}\{|X_{1,k}B_m(Z_1)\eta^{(l)}_{1,j}|^t\} \leq \mathrm{E}\{|X_{1,k}B_m(Z_1)|^{2t}\} + \mathrm{E}\{|\eta^{(l)}_{1,j}|^{2t}\} \leq C_3 t!(C_4 L^{t-2})L$$

for any positive integer $t \geq 2$ for some positive constants $C_3$ and $C_4$. By applying Bernstein's inequality with $x = C_5\sqrt{n^{-1}L\log n}$ for some suitable $C_5$, $v_i = Ln^{-2}$, and $M = O(L/n)$, we follow the standard argument and obtain

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,k}B_m(Z_i)\eta_{i,j}^{(l)}\right| \leq C_6\sqrt{\frac{L\log n}{n}} \tag{S.16}$$

uniformly in $k, m, l, j$ with probability tending to 1 for some positive constant $C_6$ depending on $C_5$.

(S.16) yields the desired result of the lemma :

$$\left\{\sum_{l=1}^{L}\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,k}B_m(Z_i)\eta_{i,j}^{(l)}\right|^2\right\}^{1/2} \leq C_6\sqrt{\frac{L^2\log n}{n}}$$

uniformly in $k, m, j$ with probability tending to 1. Hence the proof of the lemma is complete.

**Proof of Lemma 4)** We should just follow that of Lemma 2. Note that we can use the result on $\delta_n$ there as it is since it does not depend on $j$ or $l$. We should replace $\widehat{\Sigma}$, $\Sigma$, $\mathcal{S}_0$, and $s_0$ with $\widehat{\Sigma}_{-j,-j}$, $\Sigma_{-j,-j}$, $\mathcal{S}_j^{(l)}$, and $s_j^{(l)}$, respectively and then modify the definition of $\Psi(\mathcal{S}_0, 3)$ conformably.

**Proof of Proposition 2)** We should just apply the standard argument of the Lasso as in the proof of Proposition 1. Then the results follow from Lemmas 3 and 4. The details are omitted.

**Proof of Lemma 5)** Write

$$\begin{aligned}\widehat{B}_{j,k} &= \frac{1}{n}E_j^T E_k + \frac{1}{n}E_j^T \boldsymbol{W}_{-k}(\Gamma_k - \widehat{\Gamma}_k) + \frac{1}{n}(\Gamma_j - \widehat{\Gamma}_j)^T \boldsymbol{W}_{-j}^T E_k + \frac{1}{n}(\Gamma_j - \widehat{\Gamma}_j)^T \boldsymbol{W}_{-j}^T \boldsymbol{W}_{-k}(\Gamma_k - \widehat{\Gamma}_k)\\ &:= \widehat{D}_1 + \widehat{D}_2 + \widehat{D}_3 + \widehat{D}_4,\end{aligned}$$

where $\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4$ are clearly defined in the last line. We evaluate $\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4$ uniformly in $j$ and $k$. We suppress the subscripts $j$ and $k$ here.
$\widehat{D}_1$ : Exactly as in the proof of Lemma 3, we have

$$\max_{1\leq a,b\leq L} |(\widehat{D}_1 - B_{j,k})_{a,b}| \leq C_1\sqrt{\frac{L\log n}{n}} \tag{S.17}$$

uniformly in $j$ and $k$ with probability tending to 1 for some positive constant $C_1$.
$\widehat{D}_2$ and $\widehat{D}_3$ : Recall the result in Proposition 2. Then the absolute value of the $(a, b)$ element of $\widehat{D}_2$ is bounded from above by

$$n^{-1/2}\|\eta_j^{(a)}\| n^{-1/2}\|\boldsymbol{W}_{-k}(\widehat{\gamma}_k^{(b)} - \gamma_k^{(b)})\| \leq C_2(s_k^{(b)})^{1/2}\sqrt{n^{-1}L^2\log n} \tag{S.18}$$

uniformly in $a, b, j, k$ with probability tending to 1 for some positive constant $C_2$. We can treat $\widehat{D}_3$ in the same way.

$\widehat{D}_4$ : By Proposition 2, the absolute value of the $(a, b)$ element of $\widehat{D}_4$ is bounded from above by

$$n^{-1/2}\|\boldsymbol{W}_{-j}(\widehat{\gamma}_j^{(a)} - \gamma_j^{(a)})\| n^{-1/2}\|\boldsymbol{W}_{-k}(\widehat{\gamma}_k^{(b)} - \gamma_k^{(b)})\| \leq C_3(s_j^{(a)} s_k^{(b)})^{1/2} n^{-1} L^2 \log n \tag{S.19}$$

uniformly in $a, b, j, k$ with probability tending to 1 for some positive constant $C_3$.

By (S.17)-(S.19) and Assumption L(2), we have

$$L \max_{1\leq a,b\leq L} |(\widehat{B}_{j,k} - B_{j,k})_{a,b}| \to 0$$

uniformly in $j$ and $k$ with probability tending to 1. This implies the desired result

$$\|\widehat{B}_{j,k} - B_{j,k}\|_F \to 0$$

uniformly in $j$ and $k$ with probability tending to 1. Hence the proof of the lemma is complete.

**Proof of Lemma 6)** Write

$$T_j^2 = \frac{1}{n}\widehat{E}_j^T \widehat{E}_j + \widehat{\Gamma}_j^T K_j \Lambda_j = \widehat{B}_{j,j} + \widehat{A}_j,$$

where $\widehat{A}_j$ is defined as $\widehat{A}_j := \widehat{\Gamma}_j^T K_j \Lambda_j$. Suppose we have proved $\|\widehat{A}_j\|_F \to 0$ uniformly in $j$ with probability tending to 1. We will verify this convergence in probability at the end of the proof.

Write the singular value decomposition of $T_j^2$ as $T_j^2 = U_j^T \Pi_j V_j$, where $\Pi_j = \mathrm{diag}(\pi_1, \ldots, \pi_L)$. Lemma 5 and (S.1) imply that for any $x$ satisfying $\|x\| = 1$,

$$\lambda_{\min}(\Theta_{j,j}^{-1}) + o(1) \leq \|T_j^2 x\| \leq \lambda_{\max}(\Theta_{j,j}^{-1}) + o(1) \tag{S.20}$$

uniformly in $j$ with probability tending to 1. This is because $\|\widehat{A}_j x\| \leq \|\widehat{A}_j\|_F$. Recall also that $B_{j,j} = \Theta_{j,j}^{-1}$. (S.20) implies that

$$\lambda_{\min}^2(\Theta_{j,j}^{-1}) + o(1) \leq \min\{\pi_1^2, \ldots, \pi_L^2\} \leq \max\{\pi_1^2, \ldots, \pi_L^2\} \leq \lambda_{\max}^2(\Theta_{j,j}^{-1}) + o(1) \tag{S.21}$$

uniformly in $j$ with probability tending to 1. (a) follows from (S.21) and (25) since

$$\rho^2(T_j^2) = \max\{\pi_1^2, \ldots, \pi_L^2\} \quad \text{and} \quad \rho^2(T_j^{-2}) = 1/\min\{\pi_1^2, \ldots, \pi_L^2\}.$$

Next we demonstrate (b). Since

$$T_j^2 - \Theta_{j,j}^{-1} = \widehat{B}_{j,j} - \Theta_{j,j}^{-1} + \widehat{A}_j,$$

the first result follows from Lemma 5. As for the second result, notice that

$$T_j^{-2} - \Theta_{j,j}^{-1} = T_j^{-2}(\Theta_{j,j}^{-1} - T_j^2)\Theta_{j,j}. \tag{S.22}$$

The second result follows from (a), the first one, and (25).

$\|\widehat{A}_j\|_F$ : The $(a, b)$ element of $\widehat{A}_j$ is bounded from above by

$$\sum_{k\neq j} |\lambda_j^{(b)} \widehat{\gamma}_{j,k}^{(a)T} \kappa_{j,k}^{(b)}| \leq \sum_{k\neq j} \lambda_j^{(b)} \|\widehat{\gamma}_{j,k}^{(a)}\| = \lambda_j^{(b)} \mathrm{P}_1(\widehat{\gamma}_j^{(a)}).$$

Therefore

$$\|\widehat{A}_j\|_F \leq L \max_{a,b.j} \{\lambda_j^{(b)} \mathrm{P}_1(\widehat{\gamma}_j^{(a)})\} \leq C \sqrt{\frac{L^4 \log n}{n}} (\max_{a,j} s_j^{(a)})^{1/2} \to 0$$

uniformly in $j$ with probability tending to 1 for some positive constant $C$. We used Proposition 2, Assumptions S2(1) and L(2), and the fact that $\mathrm{P}_1(\gamma_j^{(a)}) \leq (s_j^{(a)})^{1/2} \|\gamma_j^{(a)}\|$. Hence the proof of the lemma is complete.

## S.2 Additional numerical studies

### S.2.1 Simulation studies

We present MSE results of our simulation studies here. We compared the oracle estimator, the original group Lasso, the adaptive group Lasso(ALasso), the group SCAD, and the de-biased group Lasso in terms of MSE defined below in (S.23). From a theoretical point of view, the group SCAD has the same asymptotic covariance matrix as the oracle estimator since the SCAD is selection consistent and a post-selection estimator. Actually the SCAD is almost the best in MSE among the original group Lasso, the adaptive group Lasso, the group SCAD, and the de-biased group Lasso. However, we should emphasize again that the de-biased group Lasso is the estimator without any variable selection and that it is used for statistical inference under the original high-dimensional model. We are not able to carry out this kind of statistical inference with the SCAD because it selects covariates.

The models and the parameters such as $n$ and $p$ are the same as in Section 4. We used also the cv.gglasso function as well as in Section 4. We implemented the group SCAD by using the R package 'grpreg' version 3.2-1 (the cv.grpreg function). It is provided by Prof. Patrick Breheny. See [2] for more details. Our weights of the adaptive group Lasso estimator are as follows:

$$w_j := \frac{1}{\max\{\|\widehat{\beta}_j\|, 0.001\}}.$$

Let $\overline{g}_j$ be an estimator of $g_j$. Then MSE and AME in tables are defined as

$$\text{MSE} := \text{the average over the repetitions of } \frac{1}{n} \sum_{i=1}^{n} f_j^2(Z_i), \tag{S.23}$$

$f_j = g_j$ or $\overline{g}_j - g_j$ for relevant $j \in \mathcal{S}_0$ and

$$\text{AMSE} := \text{the average over the repetitions of } \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n} f_j^2(Z_i), \qquad f_j = \overline{g}_j$$

for $\mathcal{S} = \{1,3,5,7,9,10,11,12\}$ (Models 1-2) and $\{1,3,5,7,9,11\}$ (Model 3). The group SCAD and the adaptive group Lasso selected almost no variable from $\mathcal{S} = \{1,3,5,7,9,10,11,12\}$ (Models 1-2) and $\{1,3,5,7,9,11\}$ (Model 3) and AMSE in the captions is that of the de-biased group Lasso.

Table S.1: MSE for Model 1 with $p = 250$ (AMSE = 0.0582)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 7.2448 | 2.3130 | 2.0411 | 2.0981 |
| oracle | 0.0758 | 0.0853 | 0.0764 | 0.0766 |
| Lasso | 0.2670 | 0.3371 | 0.2225 | 0.1698 |
| ALasso | 0.1101 | 0.2708 | 0.1829 | 0.1295 |
| SCAD | 0.0659 | 0.0916 | 0.0849 | 0.0852 |
| de-biased | 0.0933 | 0.1233 | 0.1003 | 0.1004 |

Table S.2: MSE for Model 2 with $p = 250$ (AMSE = 0.0639)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 4.4265 | 1.8147 | 2.1168 | 1.9670 |
| oracle | 0.0761 | 0.0854 | 0.0764 | 0.0767 |
| Lasso | 0.2408 | 0.2628 | 0.1524 | 0.1653 |
| ALasso | 0.0841 | 0.1458 | 0.0920 | 0.0974 |
| SCAD | 0.0668 | 0.0965 | 0.0861 | 0.0863 |
| de-biased | 0.0916 | 0.1209 | 0.0911 | 0.0962 |

Table S.3: MSE for Model 3 with $p = 250$ (AMSE = 0.0563)

| $j$ | 2 | 4 | 6 | 8 | | |
|---|---|---|---|---|---|---|
| $g_j$ | 4.4265 | 2.3130 | 2.1168 | 1.9670 | 2.0411 | 1.8147 |
| oracle | 0.0810 | 0.0922 | 0.0808 | 0.0885 | 0.0862 | 0.0813 |
| Lasso | 0.2829 | 0.3322 | 0.2262 | 0.1452 | 0.1866 | 0.2529 |
| ALasso | 0.0955 | 0.1685 | 0.1283 | 0.0911 | 0.1049 | 0.1325 |
| SCAD | 0.0723 | 0.0956 | 0.0882 | 0.0944 | 0.0871 | 0.0840 |
| de-biased | 0.1164 | 0.1413 | 0.1126 | 0.1076 | 0.1123 | 0.1314 |

Table S.4: MSE for Model 1 with $p = 350$ (AMSE = 0.0410)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 7.0290 | 2.1115 | 2.0634 | 2.1013 |
| oracle | 0.0504 | 0.0532 | 0.0570 | 0.0500 |
| Lasso | 0.1936 | 0.2512 | 0.1615 | 0.1252 |
| ALasso | 0.0926 | 0.2317 | 0.1597 | 0.1027 |
| SCAD | 0.0522 | 0.0562 | 0.0592 | 0.0570 |
| de-biased | 0.0688 | 0.0769 | 0.0696 | 0.0695 |

Table S.5: MSE for Model 2 with $p = 350$ (AMSE = 0.0445)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 4.6034 | 2.0210 | 2.0928 | 2.0379 |
| oracle | 0.0508 | 0.0533 | 0.0570 | 0.0500 |
| Lasso | 0.1751 | 0.1857 | 0.1085 | 0.1212 |
| ALasso | 0.0680 | 0.1052 | 0.0720 | 0.0720 |
| SCAD | 0.0526 | 0.0584 | 0.0593 | 0.0576 |
| de-biased | 0.0685 | 0.0721 | 0.0642 | 0.0691 |

Table S.6: MSE for Model 3 with $p = 350$ (AMSE = 0.0414)

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $g_j$ | 4.6034 | 2.1115 | 2.0928 | 2.0379 | 2.0634 | 2.0210 |
| oracle | 0.0527 | 0.0558 | 0.0599 | 0.0552 | 0.0513 | 0.0538 |
| Lasso | 0.1952 | 0.2393 | 0.1591 | 0.1024 | 0.1260 | 0.1796 |
| ALasso | 0.0760 | 0.1306 | 0.0995 | 0.0653 | 0.0782 | 0.1071 |
| SCAD | 0.0557 | 0.0603 | 0.0610 | 0.0647 | 0.0607 | 0.0562 |
| de-biased | 0.0792 | 0.0855 | 0.0742 | 0.0754 | 0.0714 | 0.0819 |

We also present the results on the other three models, Model 1', Model 2' and Model 3'. We defined them by replacing $g_j$ with $g_j/\sqrt{2}$ in Models 1-3.

Model 1'($p = 250$ and $n = 250$)

Table S.7: $H_1$ for Model 1' with $p = 250$ and $n = 250$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table S.8: $H_0$ for Model 1' with $p = 250$ and $n = 250$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.11 | 0.06 | 0.06 | 0.16 | 0.10 | 0.08 | 0.15 | 0.10 |
| $\alpha = 0.05$ | 0.06 | 0.01 | 0.03 | 0.12 | 0.06 | 0.05 | 0.07 | 0.06 |

Table S.9: MSE for Model 1' with $p = 250$ (AMSE = 0.0596)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 3.6224 | 1.1565 | 1.0206 | 1.0490 |
| oracle | 0.0758 | 0.0853 | 0.0764 | 0.0766 |
| Lasso | 0.2574 | 0.3138 | 0.2025 | 0.1526 |
| ALasso | 0.0878 | 0.2178 | 0.1406 | 0.1026 |
| SCAD | 0.0678 | 0.1171 | 0.1107 | 0.0985 |
| de-biased | 0.0873 | 0.1134 | 0.0925 | 0.0940 |

Model 2'($p = 250$ and $n = 250$)

Table S.10: $H_1$ for Model 2' with $p = 250$ and $n = 250$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table S.11: $H_0$ for Model 2' with $p = 250$ and $n = 250$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.12 | 0.10 | 0.16 | 0.18 | 0.12 | 0.08 | 0.14 | 0.12 |
| $\alpha = 0.05$ | 0.06 | 0.06 | 0.10 | 0.10 | 0.06 | 0.04 | 0.08 | 0.05 |

Table S.12: MSE for Model 2' with $p = 250$ (AMSE = 0.0641)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 2.2132 | 0.9073 | 1.0584 | 0.9835 |
| oracle | 0.0760 | 0.0853 | 0.0764 | 0.0767 |
| Lasso | 0.2195 | 0.2291 | 0.1354 | 0.1471 |
| ALasso | 0.0722 | 0.1199 | 0.0807 | 0.0829 |
| SCAD | 0.0711 | 0.1225 | 0.1023 | 0.1002 |
| de-biased | 0.0841 | 0.1102 | 0.0858 | 0.0899 |

Model 3'($p = 250$ and $n = 250$)

Table S.13: $H_1$ for Model 3' with $p = 250$ and $n = 250$

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table S.14: $H_0$ for Model 3' with $p = 250$ and $n = 250$

| $j$ | 1 | 3 | 5 | 7 | 9 | 11 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.14 | 0.07 | 0.05 | 0.20 | 0.20 | 0.14 |
| $\alpha = 0.05$ | 0.09 | 0.04 | 0.02 | 0.14 | 0.15 | 0.09 |

Table S.15: MSE for Model 3' with $p = 250$ (AMSE = 0.0575)

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $g_j$ | 2.2132 | 1.1565 | 1.0584 | 0.9835 | 1.0206 | 0.9073 |
| oracle | 0.0809 | 0.0922 | 0.0808 | 0.0885 | 0.0862 | 0.0812 |
| Lasso | 0.2615 | 0.3005 | 0.2022 | 0.1261 | 0.1614 | 0.2184 |
| ALasso | 0.0837 | 0.1459 | 0.1097 | 0.0830 | 0.0898 | 0.1080 |
| SCAD | 0.0745 | 0.1164 | 0.1078 | 0.1063 | 0.1023 | 0.1018 |
| de-biased | 0.1036 | 0.1250 | 0.1016 | 0.0982 | 0.1000 | 0.1163 |

Model 1'($p = 350$ and $n = 350$)

Table S.16: $H_1$ for Model 1' with $p = 350$ and $n = 350$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table S.17: $H_0$ for Model 1' with $p = 350$ and $n = 350$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.10 | 0.03 | 0.05 | 0.16 | 0.10 | 0.08 | 0.10 | 0.08 |
| $\alpha = 0.05$ | 0.06 | 0.02 | 0.03 | 0.10 | 0.06 | 0.04 | 0.06 | 0.05 |

Table S.18: MSE for Model 1' with $p = 350$ (AMSE = 0.0419)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 3.5145 | 1.0557 | 1.0317 | 1.0506 |
| oracle | 0.0504 | 0.0532 | 0.0570 | 0.0500 |
| Lasso | 0.1874 | 0.2380 | 0.1501 | 0.1154 |
| ALasso | 0.0702 | 0.1602 | 0.1144 | 0.0762 |
| SCAD | 0.0538 | 0.0649 | 0.0707 | 0.0657 |
| de-biased | 0.0658 | 0.0725 | 0.0658 | 0.0664 |

Model 2'($p = 350$ and $n = 350$)

Table S.19: $H_1$ for Model 2' with $p = 350$ and $n = 350$

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 |

Table S.20: $H_0$ for Model 2' with $p = 350$ and $n = 350$

| $j$ | 1 | 3 | 5 | 7 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.10 | 0.10 | 0.11 | 0.16 | 0.10 | 0.06 | 0.12 | 0.08 |
| $\alpha = 0.05$ | 0.05 | 0.06 | 0.06 | 0.10 | 0.06 | 0.04 | 0.06 | 0.06 |

Table S.21: MSE for Model 2' with $p = 350$ (AMSE = 0.0449)

| $j$ | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $g_j$ | 2.3017 | 1.0105 | 1.0464 | 1.0189 |
| oracle | 0.0506 | 0.0532 | 0.0570 | 0.0500 |
| Lasso | 0.1618 | 0.1678 | 0.0987 | 0.1120 |
| ALasso | 0.0564 | 0.0797 | 0.0606 | 0.0597 |
| SCAD | 0.0544 | 0.0718 | 0.0682 | 0.0650 |
| de-biased | 0.0647 | 0.0679 | 0.0614 | 0.0664 |

Model 3'($p = 350$ and $n = 350$)

Table S.22: $H_1$ for Model 3' with $p = 350$ and $n = 350$

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $\alpha = 0.05$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table S.23: $H_0$ for Model 3' with $p = 350$ and $n = 350$

| $j$ | 1 | 3 | 5 | 7 | 9 | 11 |
|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | 0.09 | 0.04 | 0.07 | 0.22 | 0.18 | 0.10 |
| $\alpha = 0.05$ | 0.08 | 0.03 | 0.03 | 0.16 | 0.12 | 0.06 |

Table S.24: MSE for Model 3' with $p = 350$ (AMSE = 0.0420)

| $j$ | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| $g_j$ | 2.3017 | 1.0557 | 1.0464 | 1.0189 | 1.0317 | 1.0105 |
| oracle | 0.0525 | 0.0558 | 0.0599 | 0.0552 | 0.0513 | 0.0537 |
| Lasso | 0.1839 | 0.2241 | 0.1475 | 0.0922 | 0.1127 | 0.1636 |
| ALasso | 0.0632 | 0.0980 | 0.0801 | 0.0584 | 0.0645 | 0.0839 |
| SCAD | 0.0571 | 0.0654 | 0.0689 | 0.0709 | 0.0667 | 0.0676 |
| de-biased | 0.0737 | 0.0786 | 0.0694 | 0.0705 | 0.0661 | 0.0756 |

### S.2.2 A real data application

We applied the proposed de-biased group Lasso procedure to the Boston Housing data as in e.g. [S1] and [S3]. The data set is available in the R package 'MASS.' See also [S2] about the data set. The data set has 14 variables, crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv, and 506 samples. The details of these variables are given at the end of this section. We augmented the data set by adding some artificial variables.

In this study, we followed [S1] and [S3] and took $Y$ = medv and lstat as the index variable. Note that [S1] does not deal with high-dimensional models. As for lstat, we defined $Z$ as $Z = F(\text{lstat})$, where $F(\cdot)$ is the distribution function of 2× the $\chi^2$ distribution with d.f. 6. We did this transformation to make the distribution of $Z$ close to that of the uniform distribution on $[0, 1]$. Note that [S1] and [S3] included only part of the original variables e.g. crim, rm, tax, and ptratio in their models. We removed only a dummy variable chas since it does seem to be significant in our preliminary analysis. The conditional number of the covariance matrix of 11 original variables exceeds 100. This setup is unfavorable to any data analysis procedure. The conditional number of the covariance matrix of only crim, rm, tax, and ptratio is about 14.

In this section, we present two results : the one with 11 original variables and 89 augmented variables in Table S.25 and the one with only 11 original variables in Table S.26.

We explain our augmented model. Let $q$ be the number of the original variables ($q = 11$). Then our augmented model is

$$Y = g_0(Z) + \sum_{j=1}^{q} g_j(Z)X_j + \sum_{j=q+1}^{p} g_j(Z)X_j + \epsilon. \tag{S.24}$$

First we standardized the $q$ original variables so that they have mean 0 and variance 1 and got $X_1, \ldots, X_q$. The details of the artificial variables are as follows:

$$X'_{j+11} = 0.25X_j + 0.75R_j, \quad j = 1, \ldots, q,$$

where $R_j$, $j = 1, \ldots, q$, are i.i.d. $\mathrm{N}(0, 1)$ random variables. Then we standardized $X'_{q+1}, \ldots, X'_{2q}$ as well and defined $X_{q+1}, \ldots, X_{2q}$ from them. $X'_{2q+1}, \ldots, X'_p$ are i.i.d normal random variables and we also standardized them to define $X_{2q+1}, \ldots, X_p$.

In the tables, $||\widehat{b}_j||^2$ and $||\widetilde{\beta}_j||^2$ are from the de-biased Lasso and the SCAD, respectively. We computed p-values in the tables in a similar way to the critical values in Section 4 by using Theorem 1. We tried $p = 100$ with $L = 5$ and the quadratic spline basis. The results of 24 larger $||\widehat{b}_j||^2$ are given in Table S.25. In [S3], they included only four original variables (rm, crim, tax, ptratio) and straightforward comparisons are very difficult.

If we compute all $\widehat{b}_j$ for a large $p$, it will take a very long time. Therefore some kind of screening that chooses rather many covariates and does not miss relevant variables may be necessary in practical situations.

In the two tables, the de-biased Lasso and the SCAD show different behaviors. The two tables also show different results. The original Lasso selected only two variables, rm and ptratio, in either model. This may be due to the large conditional number larger than 100 among the 11 original variables. Even the SCAD and the Lasso may have difficulty dealing with such highly correlated data sets. As for the augmented variables, some have p-values less than 0.05. But most of the augmented variables have larger p-values.

Table S.25: The model with 11 original variates and 89 augmented variables

| Variable | black | zn | rm | rad | tax | dis |
|---|---|---|---|---|---|---|
| $\|\|\widehat{b}_j\|\|^2/\text{Var}(Y)$ | 0.120 | 0.116 | 0.114 | 0.074 | 0.049 | 0.049 |
| $\|\|\widetilde{\beta}_j\|\|^2/\text{Var}(Y)$ | 0.005 | 0.000 | 0.082 | 0.112 | 0.000 | 0.062 |
| p-value | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Variable | crim | 14 | indus | ptratio | nox | 77 |
| $\|\|\widehat{b}_j\|\|^2/\text{Var}(Y)$ | 0.028 | 0.026 | 0.024 | 0.024 | 0.017 | 0.009 |
| $\|\|\widetilde{\beta}_j\|\|^2/\text{Var}(Y)$ | 0.000 | 0.000 | 0.000 | 0.059 | 0.06 | 0.000 |
| p-value | 0.015 | 0.000 | 0.002 | 0.000 | 0.120 | 0.016 |
| Variable | 42 | 21 | 37 | 80 | 88 | 59 |
| $\|\|\widehat{b}_j\|\|^2/\text{Var}(Y)$ | 0.009 | 0.008 | 0.007 | 0.006 | 0.006 | 0.006 |
| $\|\|\widetilde{\beta}_j\|\|^2/\text{Var}(Y)$ | 0.001 | 0.000 | 0.000 | 0.001 | 0.000 | 0.001 |
| p-value | 0.011 | 0.015 | 0.034 | 0.055 | 0.056 | 0.062 |
| Variable | 97 | 74 | 24 | 53 | 65 | 64 |
| $\|\|\widehat{b}_j\|\|^2/\text{Var}(Y)$ | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 | 0.006 |
| $\|\|\widetilde{\beta}_j\|\|^2/\text{Var}(Y)$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| p-value | 0.049 | 0.052 | 0.052 | 0.064 | 0.084 | 0.086 |

Table S.26: The model with only 11 original variates

| Variable | zn | black | rm | rad | tax | dis |
|---|---|---|---|---|---|---|
| $\|\|\widehat{b}_j\|\|^2/\text{Var}(Y)$ | 0.128 | 0.117 | 0.090 | 0.075 | 0.050 | 0.049 |
| $\|\|\widetilde{\beta}_j\|\|^2/\text{Var}(Y)$ | 0.000 | 0.005 | 0.073 | 0.264 | 0.057 | 0.115 |
| p-value | 0.000 | 0.092 | 0.000 | 0.000 | 0.002 | 0.000 |
| Variable | ptratio | crim | indus | nox | age | NA |
| $\|\|\widehat{b}_j\|\|^2/\text{Var}(Y)$ | 0.030 | 0.030 | 0.021 | 0.020 | 0.006 | NA |
| $\|\|\widetilde{\beta}_j\|\|^2/\text{Var}(Y)$ | 0.046 | 0.117 | 0.043 | 0.028 | 0.000 | NA |
| p-value | 0.000 | 0.015 | 0.026 | 0.097 | 0.512 | NA |

$||\widehat{b}_j||^2$ and p-value are from the de-biased group Lasso and $||\widetilde{\beta}_j||^2$ is from the group SCAD.

We reproduced the details of 14 variables from the R documentation of the R package ‘MASS.’

```
crim    : per capita crime rate by town(We took the logarithm in this section.)
zn      : proportion of residential land zoned for lots over 25,000 sq.ft
indus   : proportion of non-retail business acres per town
chas    : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
          This is not used in our model.
nox     : nitric oxides concentration (parts per 110 million)
rm      : average number of rooms per dwelling
age     : proportion of owner-occupied units built prior to 1940
dis     : weighted distances to five Boston employment centres
rad     : index of accessibility to radial highways
tax     : full-value property-tax rate per USD 10,000
ptratio : pupil-teacher ratio by town
black   : 1000(B - 0.63)^2 where B is the proportion of blacks by town
lstat   : lower status of the population
medv    : median value of owner-occupied homes in USD 1000's
```

## References

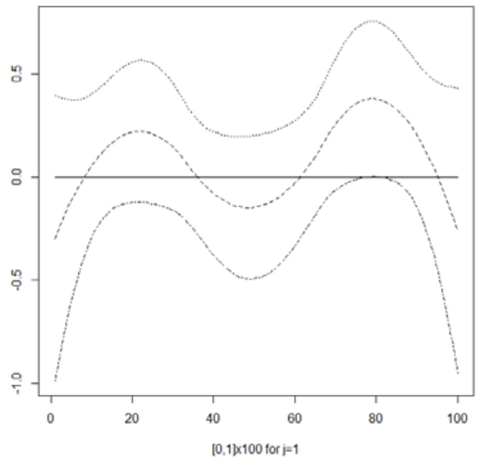
[S1] Z. Cai and X. Xu. Nonparametric quantile estimators for dynamic smooth coefficient models. *J. Amer. Statist. Assoc.*, 103:1595–1608, 2008.

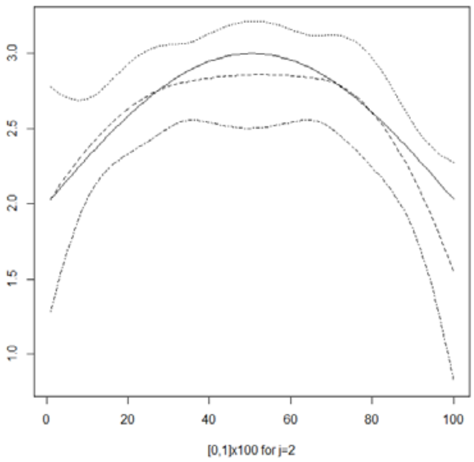
[S2] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Economics Managements*, 5:81-102, 1978.

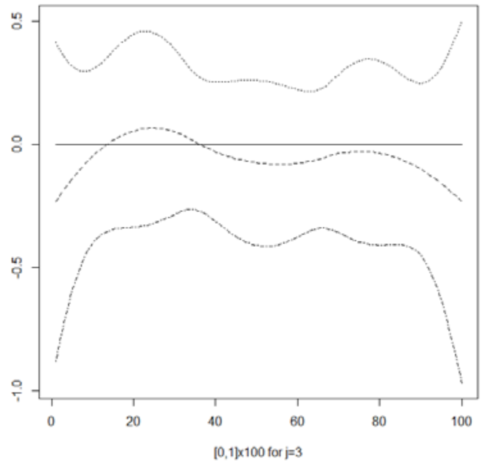
[S3] Y. Tang, X. Song, H. J. Wang, and Z. Zhu. Variable selection in high-dimensional quantile varying coefficient miodels. *J. Multivar. Anal.*, 122:115-132, 2013.

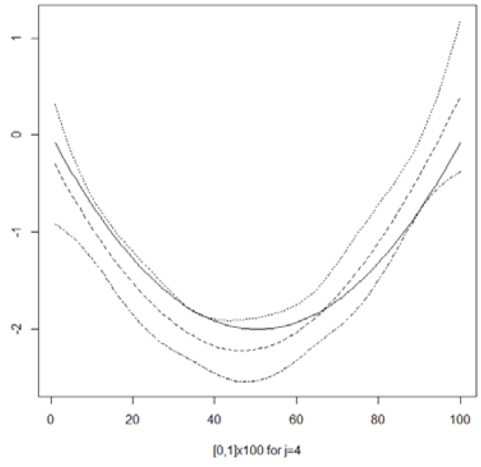
### S.2.3 Confidence bands for $g_j$

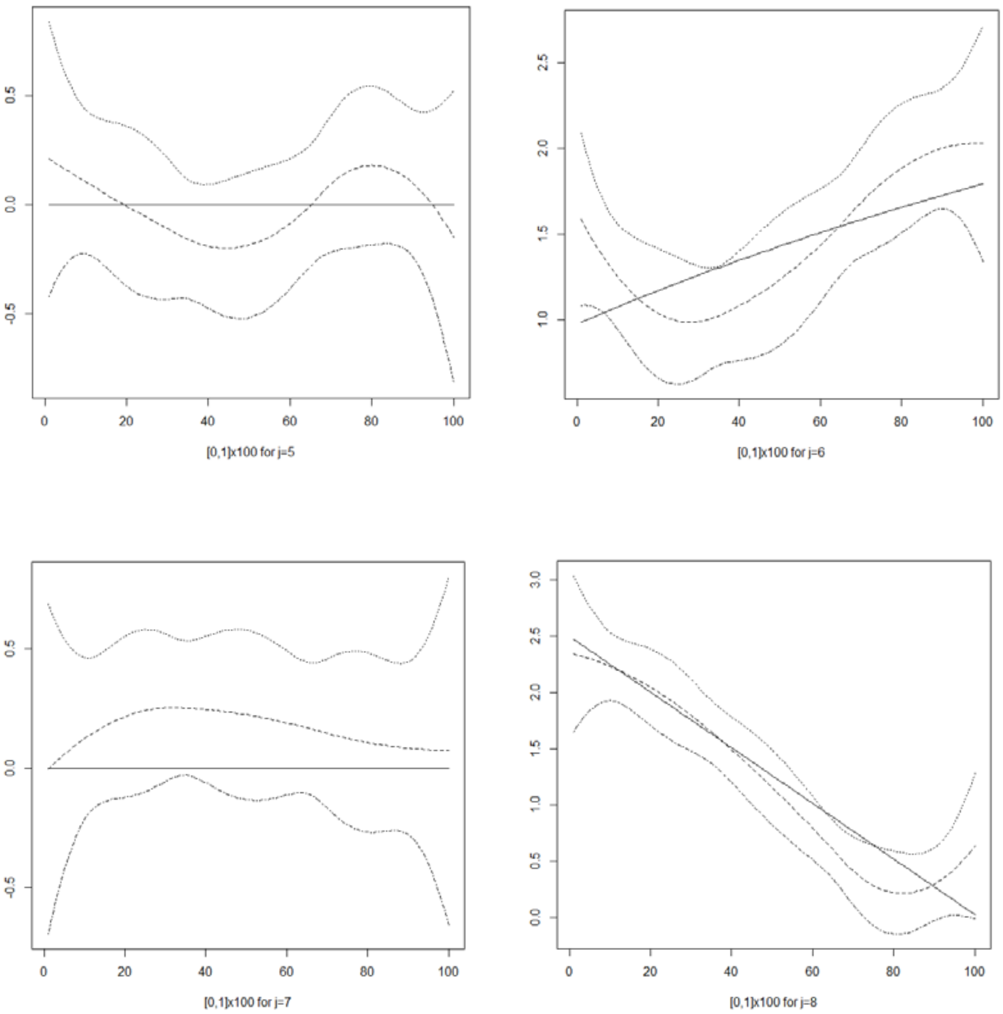
We present 8 figures of 95% confidence bands for $g_j$, j=1,⋯,8, and they are based on Theorem 1. We took one simulated sample for Model 1 with $p=n=350$. Real and broken lines represent ture $g_j$ and estimated $g_j$, respectively. The other two lines are lower and upper bands for $g_j(t)$, not simultaneous bands on [0,1]. The broken lines look sufficiently close to the real lines and the real lines are almost between the lower and upper bands. Therefore these figures imply our procedure is very promising.










0.5
0.0
-0.5
0
20
40
60
80
100
[0,1]x100 for j=5
2.5
2.0
1.5
1.0
0
20
40
60
80
100
[0,1]x100 for j=6
0.5
0.0
-0.5
0
20
40
60
80
100
[0,1]x100 for j=7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
[0,1]x100 for j=8