Annals of the Institute of Statistical Mathematics (2021) 73:3–29
https://doi.org/10.1007/s10463-019-00740-4

The de-biased group Lasso estimation for varying coefficient models

Toshio Honda[1]

Received: 29 November 2018 / Revised: 8 August 2019 / Published online: 9 November 2019

Abstract
There has been much attention on the de-biased or de-sparsified Lasso. The Lasso is very useful in high-dimensional settings. However, it is well known that the Lasso produces biased estimators. Therefore, several authors 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 needs desirable estimators of high-dimensional precision matrices. 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. We apply the de-biased group Lasso to varying coefficient models and examine the theoretical properties and the effects of approximation errors involved in nonparametric regression. The results of numerical studies 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:

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

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-019-00740-4) contains supplementary material, which is available to authorized users.

✉ Toshio Honda
t.honda@r.hit-u.ac.jp

[1] Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan

where $(Y_i, \underline{X}_i, Z_i)$, $i = 1, \ldots, n$, are i.i.d. observations, Y_i is a dependent variable, $\underline{X}_i = (X_{i,1}, \ldots, X_{i,p})^T \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}$ are random covariates, and an unobserved error ϵ_i follows the normal distribution with mean zero and variance σ_ϵ^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 $g_j(Z_i)$ $j = 1, \ldots, p$, are unknown smooth functions on $[0, 1]$ to be specified later in Sect. 3. The varying coefficient model is one of the most popular structured nonparametric regression models. For example, see Fan and Zhang (2008) 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 is much smaller than p and we need some variable selection procedures for high-dimensional datasets like the Lasso (e.g. Tibshirani 1996; Bickel et al. 2009), the SCAD (e.g. Fan and Li 2001), feature screening procedures based on marginal models or some index between the dependent variable and individual covariates (e.g. Fan and Song 2010), and forward variable selection procedures (e.g. Wang 2009; Ing and Lai 2011). Liu et al. (2015) 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 Zou (2006), Yuan and Lin (2006), and Lounici et al. (2011). There are too many papers on high-dimensional issues to mention, and we just name a few books for recent developments, Bühlmann and van de Geer (2011), Hastie et al. (2015), and van de Geer (2016).

Several authors considered ultra-high-dimensional or high-dimensional varying coefficient models by employing the group Lasso (e.g. Wei et al. 2011), the group SCAD (e.g. Cheng et al. 2014), feature screening procedures based on marginal models and so on (e.g. Fan et al. 2014; Liu et al. 2014), and forward variable selection procedures (e.g. Cheng et al. 2016). In Honda and Härdle (2014) and Honda and Yabe (2017), 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 sets the coefficient to be 0 if the covariate is not selected. Statistical inference under the original model is impossible for the SCAD.

Several authors (Zhang and Zhang 2014; Javanmard and Montanari 2014; van de Geer 2014) 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. Mitra and Zhang 2016; Stucky and van de Geer 2018). 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 set-up 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 Mitra and Zhang (2016) and Stucky and van de Geer (2018) 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 Sect. 2, we describe the de-biased group Lasso procedure for varying coefficient models. Then, we present our assumptions and main theoretical results in Sect. 3. Simulation study results are presented in Sect. 4. The results suggest that the proposed de-biased group Lasso will work well. Additional numerical results are given in Supplement. We prove the main theoretical results in Sect. 5. The technical proofs are also relegated to 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. $C, C_1, C_2, \ldots$, 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 $\overline{\mathcal{S}}$ and $|\mathcal{S}|$ stand for the complement and the number of the elements of an index set $\mathcal{S} \subset \{1, \ldots, p\}$, respectively. When we have two random vectors U and V, $U|V$ stands for the conditional distribution of U on V. And $N(\mu, \sigma^2)$ means the normal distribution with mean μ

and variance σ^2 and we write $U \sim N(\mu, \sigma^2)$ when U follows the normal distribution with mean μ and variance σ^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 $\widehat{\beta}$. Then, we need some desirable estimator of the precision matrix of Σ in Assumption S1 below and we denote the estimator by $\widehat{\Theta}$. We present $\widehat{\Theta}$ after we define $\widehat{\beta}$ and $\widehat{\boldsymbol{b}}$.

• *Regression spline model* We explain our regression spline model for (1). We denote the L-dimensional equispaced B-spline basis on $[0,1]$ by $B(z) = (B_1(z), \ldots, B_L(z))^T$ 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 in Sect. 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

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

where $\otimes$ is the Kronecker product, and

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

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

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

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:

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

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

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

Assumption S1

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

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

Note that $\Sigma = n^{-1}E(\boldsymbol{W}^T\boldsymbol{W})$ and we usually denote the inverse of Σ by Θ, not Σ^{-1}, as in the literature on high-dimensional precision matrices. The sample version of Σ 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 set-ups, and we denote our estimator of the inverse Θ by $\widehat{\Theta}$. We define an $n \times (p-1)L$ matrix $\boldsymbol{W}_{-j}$ by removing W_j from $\boldsymbol{W} = (W_1, \ldots, W_p)$. We consider regression of W_j to $\boldsymbol{W}_{-j}$ when we construct our $\widehat{\Theta}$.

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

$$\widehat{\beta} = (\widehat{\beta}_1^T, \ldots, \widehat{\beta}_p^T)^T := \underset{\beta \in \mathbb{R}^{pL}}{\operatorname{argmin}} \left\{ \frac{1}{n}\|Y - \boldsymbol{W}\beta\|^2 + 2\lambda_0 P_1(\beta) \right\}, \quad (6)$$

where $\beta = (\beta_1^T, \ldots, \beta_p^T)^T$ with $\beta_j \in \mathbb{R}^L$ for $j = 1, \ldots, p$, λ_0 is a suitably chosen tuning parameter, and $P_1(\beta) := \sum_{j=1}^p \|\beta_j\|$. We also use this $P_1(\cdot)$ 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 $\widehat{\beta}$ yields

$$-\frac{1}{n}\boldsymbol{W}^T(Y - \boldsymbol{W}\widehat{\beta}) + \lambda_0\kappa_0 = 0 \in \mathbb{R}^{pL}, \quad (7)$$

where $\kappa_0 = (\kappa_{0,1}^T, \ldots, \kappa_{0,p}^T)^T$ with $\kappa_{0,j} \in \mathbb{R}^L$ for $j = 1, \ldots, p$, $\|\kappa_{0,j}\| \leq 1$ for $j = 1, \ldots, p$, 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 $\widehat{\beta}$ is a biased estimator due to the L_1 penalty as we mentioned in Sect. 1. Thus, by constructing $\widehat{\Theta}$ such that $\widehat{\Theta}\widehat{\Sigma}$ is sufficiently close to I_{pL}, we define our de-biased group Lasso estimator $\widehat{\boldsymbol{b}} = (\widehat{b}_1^T, \ldots, \widehat{b}_p^T)^T \in \mathbb{R}^{pL}$ with $\widehat{b}_j \in \mathbb{R}^L$ for $j = 1, \ldots, p$ for the varying coefficient model (1) and (5) as

$$\begin{aligned}
\widehat{\boldsymbol{b}} &:= \widehat{\beta} + \widehat{\Theta}\lambda_0\kappa_0 = \widehat{\beta} + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T(Y - \boldsymbol{W}\widehat{\beta}) \\
&= \widehat{\beta} + \widehat{\Theta}\widehat{\Sigma}(\beta_0 - \widehat{\beta}) + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T(r + \epsilon) \\
&= \beta_0 + \frac{1}{n}(\widehat{\Theta}\widehat{\Sigma} - I_{pL})(\beta_0 - \widehat{\beta}) + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T(r + \epsilon) \\
&= \beta_0 + \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T\epsilon - \Delta_1 + \Delta_2, \qquad (8)
\end{aligned}$$

where we used (7) in the first line,

$$\begin{aligned}
\Delta_1 &= (\Delta_{1,1}^T, \ldots, \Delta_{1,p}^T)^T := \frac{1}{n}(\widehat{\Theta}\widehat{\Sigma} - I_{pL})(\widehat{\beta} - \beta_0) \in \mathbb{R}^{pL}, \\
\Delta_2 &= (\Delta_{2,1}^T, \ldots, \Delta_{2,p}^T)^T := \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T r \in \mathbb{R}^{pL}, \\
\Delta_1 &= \begin{pmatrix} \Delta_{1,1} \\ \vdots \\ \Delta_{1,p} \end{pmatrix} := \frac{1}{n}(\widehat{\Theta}\widehat{\Sigma} - I_{pL})(\widehat{\beta} - \beta_0) \in \mathbb{R}^{pL}, \\
\Delta_2 &= \begin{pmatrix} \Delta_{2,1} \\ \vdots \\ \Delta_{2,p} \end{pmatrix} := \frac{1}{n}\widehat{\Theta}\boldsymbol{W}^T r \in \mathbb{R}^{pL},
\end{aligned}$$

and $\Delta_{1,j} \in \mathbb{R}^L$ and $\Delta_{2,j} \in \mathbb{R}^L$ for $j = 1, \ldots, p$. We will prove that Δ_1 and Δ_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 Sect. 3.

The evaluation of Δ_2 requires more smoothness of the coefficient functions $g_j(z)$ than usual as in Assumption G in Sect. 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 $X_{i,1}, \ldots, X_{i,p}$, 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 van de Geer (2014) to the current group structure. Note that our construction is different from those of Mitra and Zhang (2016) and Stucky and van de Geer (2018) 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 notations 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 $\Sigma_{-j,-j}$:

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

and $\Sigma_{-j,j} := \Sigma_{j,-j}^T$. Note that they can be defined also from Σ as its submatrices. Furthermore, we define a $(p-1)L \times L$ matrix Γ_j as $\Gamma_j := \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$ and write $\Gamma_j = (\gamma_j^{(1)}, \ldots, \gamma_j^{(L)})$, where $\gamma_j^{(l)} \in \mathbb{R}^{(p-1)L}$ for $l = 1, \ldots, L$. We need to estimate this Γ_j to define $\widehat{\Theta}$. In this paper, we estimate $\Gamma_j = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j} = (\gamma_l^{(1)}, \ldots, \gamma_l^{(L)})$ columnwise by employing the group Lasso differently from Stucky and van de Geer (2018). 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 $\eta_j^{(l)} \in \mathbb{R}^n$, $j = 1, \ldots, L$, as

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}_j^{(l)} &= (\widehat{\gamma}_{j,1}^{(l)T},\ldots,\widehat{\gamma}_{j,j-1}^{(l)T},\widehat{\gamma}_{j,j+1}^{(l)T},\ldots,\widehat{\gamma}_{j,p}^{(l)T})^T\\ &:= \underset{\gamma\in\mathbb{R}^{(p-1)L}}{\operatorname{argmin}}\left\{\frac{1}{n}\|W_j^{(l)} - \boldsymbol{W}_{-j}\gamma\|^2 + 2\lambda_j^{(l)}P_1(\gamma)\right\},\end{aligned} \tag{13}$$

where $P_1(\gamma)$ is defined as in (6), $\widehat{\gamma}_{j,k}^{(l)} \in \mathbb{R}^L$ for $k \neq j$, $\gamma = (\gamma_1^T,\ldots,\gamma_{j-1}^T,\gamma_{j+1}^T,\ldots,\gamma_p^T)^T$ 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 $\widehat{\boldsymbol{\gamma}}_j^{(l)}$ in Proposition 2 in Sect. 3.

As in (7), we have

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

where $\kappa_j^{(l)} = (\kappa_{j,1}^{(l)T},\ldots,\kappa_{j,j-1}^{(l)T},\kappa_{j,j+1}^{(l)T},\ldots,\kappa_{j,p}^{(l)T})^T$ 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 $\widehat{\boldsymbol{\gamma}}_j^{(l)}, \kappa_{j,k}^{(l)}$, 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

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

and write

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

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

We are ready to define $\widehat{\Theta}$ by adapting the idea of van de Geer (2014) to the current set-up. 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 $T_j^2 := n^{-1}W_j^T \widehat{E}_j$. Compare (11) and (19), too.

$$\begin{aligned}\widehat{\Sigma}\widehat{\Theta}^T &= \frac{1}{n}\boldsymbol{W}^T(\boldsymbol{W}\widehat{\Theta}^T) = \frac{1}{n}\boldsymbol{W}^T(\widehat{E}_1, \ldots, \widehat{E}_p)\mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2}) \\ &= \begin{pmatrix} T_1^2 & K_{2,1}\Lambda_2 & K_{3,1}\Lambda_3 & \cdots & K_{p,1}\Lambda_p \\ K_{1,2}\Lambda_1 & T_2^2 & K_{3,2}\Lambda_3 & \cdots & K_{p,2}\Lambda_p \\ K_{1,3}\Lambda_1 & K_{2,3}\Lambda_2 & T_3^2 & \cdots & K_{p,3}\Lambda_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{1,p}\Lambda_1 & K_{2,p}\Lambda_2 & K_{3,p}\Lambda_3 & \cdots & T_p^2 \end{pmatrix} \mathrm{diag}(T_1^{-2}, \ldots, T_p^{-2}) \end{aligned} \tag{19}$$

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 Δ_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 Sect. 5. See also (12) about this definition of T_j^2.

In Sect. 4, we chose λ_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 Stucky and van de Geer (2018), the authors considered fixed design regression models and estimated all the columns of Γ_j simultaneously in a single Lasso-type penalized regression. On the other hand, we estimate Γ_j columnwise and we can apply the standard theory and also employ the standard R package to get our estimator of Γ_j. We can define another estimator of Θ just formally even if we estimate Γ_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 $\widehat{\beta}$ and $\widehat{\boldsymbol{\gamma}}_j^{(l)}$ in Propositions 1 and 2 together with technical assumptions. Then, we evaluate Δ_1 and Δ_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 Sect. 5. Theorem 1 immediately follows from those propositions. Propositions 1 and 2 will be proved in 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 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:

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

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

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

where $\mu_X(Z_i) = (\mu_{X,2}(Z_i), \ldots, \mu_{X,p}(Z_i))^T$ 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) $E(X_{i,j}) = 0$, $j = 2, \ldots, p$. Besides, $\|\mu_{X,j}\|_\infty < C_1$ for $j = 2, \ldots, p$ 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 $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and ϵ_i is independent of $(\underline{X}_i, Z_i)$.
(2) There is a constant σ^2 independent of Z_i such that $E\{\exp(\alpha^T \underline{\breve{X}}_i) | Z_i\} \le \exp(\|\alpha\|^2\sigma^2/2)$ 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 van de Geer (2014) and Javanmard and Montanari (2018). In particular, the arguments in Javanmard and Montanari (2018) 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 Caner and Kock (2018), 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 $\Omega_B = \int_0^1 B(z)B^T(z)\mathrm{d}z$. See, for example, Huang et al. (2004) 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 Δ_2. If we take $d_j = \|g_j\|_\infty + \|g_j'\|_\infty + \|g_j''\|_\infty + \|g_j^{(3)}\|_\infty$ and some suitable β_{0j}, this $\{d_j\}$ satisfies the first one in (23). See, for example, Corollary 6.26 of Schumaker (2007). 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* $\widehat{\beta}$ The theoretical results on the Lasso crucially depend 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*

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

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

We also use $P_\infty(\cdot)$ for vectors of lower dimension as we use $P_1(\cdot)$.

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

$$\Psi(\mathcal{S}, m) := \{\beta \in \mathbb{R}^{pL} \mid P_1(\beta_{\overline{\mathcal{S}}}) \leq m P_1(\beta_{\mathcal{S}}) \text{ and } \beta \neq 0\},$$

where $\beta_{\mathcal{S}}$ consists of $\{\beta_j\}_{j\in\mathcal{S}}$, $\beta_{\overline{\mathcal{S}}}$ consists of $\{\beta_j\}_{j\in\overline{\mathcal{S}}}$, and $P_1(\cdot)$ 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 nonnegative $(pL) \times (pL)$ matrix Ω as

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

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.

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

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

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,*

$$\frac{1}{n}\|\boldsymbol{W}(\widehat{\beta} - \beta_0)\|^2 \leq 18\frac{\lambda_0^2 s_0}{\phi^2_\Sigma(\mathcal{S}_0, 3)} \quad \text{and} \quad P_1(\widehat{\beta} - \beta_0) \leq 24\frac{\lambda_0 s_0}{\phi^2_\Sigma(\mathcal{S}_0, 3)}.$$

We will prove this proposition in Supplement including the case where we have some prior knowledge on $\mathcal{S}_0$. Note that C in the definition of λ_0 is from Lemma 1. We will follow the proof in Caner and Kock (2018), and we can also deal with the weighted group Lasso as in Caner and Kock (2018) with just conformable changes. Note that Caner and Kock (2018) considered the adaptive Lasso for linear regression models. We did not present the adaptively weighted Lasso version since the notation is very complicated in the current set-up 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 Fan et al. (2014), it is not biased and we do not 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 $\widehat{\boldsymbol{\gamma}}_j^{(l)}$ for $\widehat{\Theta}$* We consider the properties of another group Lasso estimator $\widehat{\boldsymbol{\gamma}}_j^{(l)}$ defined in (13). We deal with the deviation condition and the RE condition in Lemmas 3 and 4, respectively. Then, Proposition 2 about the group Lasso estimator $\widehat{\boldsymbol{\gamma}}_j^{(l)}$ in (13) follows almost automatically from them.

We define the active index set $\mathcal{S}_j^{(l)} \subset \{1, \ldots, j-1, j+1, \ldots, p\}$ of $\gamma_j^{(l)}$ in almost the same way as $\mathcal{S}_0$ of β_0 and let $s_j^{(l)} := |\mathcal{S}_j^{(l)}|$. We assume $\mathcal{S}_j^{(l)}$ 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) $\|\gamma_j^{(l)}\| \leq C_1$ 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

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

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 $\gamma_j^{(l)}$ in (13). Then, we write $\eta_j^{(l)} = (\eta_{1,j}^{(l)}, \ldots, \eta_{n,j}^{(l)})^T \in \mathbb{R}^n$. Since $\Sigma_{-j,j} - \Sigma_{-j,-j}\Gamma_j = 0$ and our observations are i.i.d., we have

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

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

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

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

uniformly in l $(1 \leq l \leq L)$ and j $(1 \leq j \leq p)$. 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 $(1 \leq l \leq L)$ and j $(1 \leq j \leq p)$,

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

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

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*

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

uniformly in l $(1 \leq l \leq L)$ and j $(1 \leq j \leq p)$ 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*

$$\frac{1}{n}\|\boldsymbol{W}_{-j}(\widehat{\boldsymbol{\gamma}}_j^{(l)} - \gamma_j^{(l)})\|^2 \leq 18\frac{\lambda_j^{(l)2}s_j^{(l)}}{\lambda_{\min}(\Sigma)} \quad \text{and} \quad P_1(\widehat{\boldsymbol{\gamma}}_j^{(l)} - \gamma_j^{(l)}) \leq 24\frac{\lambda_j^{(l)}s_j^{(l)}}{\lambda_{\min}(\Sigma)}$$

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 λ_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 $(1 \le j \le p)$ *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*

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

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

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 $\widehat{\eta}_j^{(l)} := W_j - \boldsymbol{W}_{-j}^T \widehat{\boldsymbol{\gamma}}_j^{(l)} \in \mathbb{R}^n$ 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

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

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 Σ and Θ, 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* $\{j_1, \ldots, j_m\} \subset \{1, \ldots, p\}$, *we define a symmetric matrix* Δ *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 $\{j_1, \ldots, j_m\}$ *with probability tending to 1.*

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

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

$$\widehat{\boldsymbol{b}} - \beta_0 = \frac{1}{n}\widehat{\Theta}\boldsymbol{W}\epsilon - \Delta_1 + \Delta_2$$

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 $n^{-1/2}\widehat{\Theta}\boldsymbol{W}^T\epsilon \mid \{\underline{X}_i, Z_i\}_{i=1}^n \sim N(0, \sigma_\epsilon^2\widehat{\Omega})$ and $\widehat{\Omega}$ converges in probability to Θ 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 $\int B'(z)(B'(z))^T \mathrm{d}z = I_L$. 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 Δ_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

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

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. See Subsection S.2.3 in 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 $b_{s,j}^{(l)}$ is part of $\gamma_j^{(l)}$. If $\sum_{s \in \mathcal{S}_j^{(l)}} \|b_{s,j}^{(l)}\| < C$ or $s_j^{(l)} < C$ 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).

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

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, $\sum_{m=1}^{L} B_m^2(Z_1) < C_3 L$ 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

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

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

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

uniformly in j with probability tending to 1 for some positive constant C, this is small enough. Hence, we have only to evaluate $n^{-1}r^T\eta_j^{(l)}$. Recalling $r_i = \sum_{j\in\mathcal{S}_0} X_{i,j}(g_j(Z_i) - B^T(Z_i)\beta_{0j})$ and $\eta_j^{(l)} = W_j^{(l)} - \boldsymbol{W}_{-j}\gamma_j^{(l)}$, we conjecture that $n^{-1}r^T\eta_j^{(l)}$ 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 $j = 1,\ldots,12$ in Models 1–3 defined below. We also present some more simulation results and a real data application in Section S.2 in 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 Yang and Zou (2017) 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 did not optimize $\widehat{\boldsymbol{b}}$ with respect to λ_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 $E\{X_{i,2}\} = 0$ and $E\{X_{i,2}^2\} = 1$ and Z_i and $\{X_{i,j}\}_{j=2}^{p}$ are mutually independent. As for the error term, we took $\epsilon_i \sim N(0, 3)$. We tried two cases, $(L, p, n,$ 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

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

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

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

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 $j = 1, \ldots, 12$ in Models 1–3. We computed the critical values from the result that $\sqrt{n}(\widehat{b}_j - \beta_{0j})$ is approximately distributed as $N(0, \widehat{\Omega}_{j,j})$ 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 values. The nominal significance levels are 0.05 and 0.10.

In Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 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 six more cases where every $g_j(z)$ is replaced with $g_j(z)/\sqrt{2}$. There is no significant differences and the

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

results of the six cases are presented in 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 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 Supplement.

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

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

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

$$B_{j,j} = \Sigma_{j,j} - \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} = \Theta_{j,j}^{-1} \quad \text{and}$$

$$B_{j,k} = E\left\{ \begin{pmatrix} \eta_{1,j}^{(1)} \\ \vdots \\ \eta_{1,j}^{(L)} \end{pmatrix} (\eta_{1,k}^{(1)}, \ldots, \eta_{1,k}^{(L)}) \right\}. \tag{33}$$

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

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

and

$$|\lambda_j^{(l)} \sum_{k \ne j} \kappa_{j,k}^{(l)T} (\widehat{\beta}_k - \beta_{0k})| \le \max_{a,b} \lambda_a^{(b)} P_1(\widehat{\boldsymbol{\beta}} - \beta_0),$$

we have uniformly in j,

$$\|\Delta_{1,j}\| \le \max_{a,b} \lambda_a^{(b)} \rho(T_j^{-2}) L^{1/2} P_1(\widehat{\boldsymbol{\beta}} - \beta_0). \tag{34}$$

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 $P_1(\widehat{\beta} - \beta_0)$ 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 λ_0 in Proposition 1. Hence, the proof of the proposition is complete. □

Proof of Proposition 4 Write

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

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

$$P_\infty(n^{-1}\boldsymbol{W}^T r)(P_1(\widehat{\boldsymbol{\gamma}}_j^{(l)}) + 1) \le C_1 P_\infty(n^{-1}\boldsymbol{W}^T r)(s_j^{(l)})^{1/2}(\|\boldsymbol{\gamma}_j^{(l)}\| + 1) \tag{36}$$

uniformly in l and j with probability tending to 1 for some positive constant C_1. Note that the difference between $P_1(\widehat{\boldsymbol{\gamma}}_j^{(l)})$ and $P_1(\boldsymbol{\gamma}_j^{(l)})$ is negligible by Proposition 2 and that $P_1(\boldsymbol{\gamma}_j^{(l)}) \le (s_j^{(l)})^{1/2}\|\boldsymbol{\gamma}_j^{(l)}\|$.

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

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

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

(37) and Lemma 1 yield the desired result. 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 $\{j_1, \ldots, j_m\}$ 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

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

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 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

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

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

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

In addition,

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

Then (44) and (A-74) in Greene (2012) 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 $U_1, \ldots, U_p$ satisfying

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

We define a set of $\Theta_{1,1}$, $\Theta_{1,2}$, $\Theta_{2,2}$, $B_{1,2}$ for this Σ 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 Greene (2012), we have

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

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 Σ^{-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. □

Acknowledgements The author appreciates comments from the AE and two reviewers very much. He also thanks Akira Shinkyu for research assistance. This research is supported by JSPS KAKENHI Grant Number JP 16K05268.

References

Bickel, P. J., Ritov, Y., Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, *37*, 1705–1732.

Bühlmann, P., van de Geer, S. (2011). *Statistics for high-dimensional data: Methods theory and applications*. New York: Springer.

Caner, M., Kock, A. B. (2018). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, *203*, 143–168.

Cheng, M.-Y., Honda, T., Li, J., Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Annals of Statistics*, *42*, 1819–1849.

Cheng, M.-Y., Honda, T., Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, *111*, 1201–1221.
Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.
Fan, J., Ma, Y., Dai, W. (2014). Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, *109*, 1270–1284.
Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, *38*, 3567–3604.
Fan, J., Xue, L., Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, *42*, 819–849.
Fan, J., Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, *1*, 179–195.
Greene, W. H. (2012). *Econometric analysis* 7th ed. Harlow: Pearson Education.
Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical learning with sparsity*. Boca Raton: CRC Press.
Honda, T., Härdle, W. K. (2014). Variable selection in Cox regression models with varying coefficients. *Journal of Statistical Planning and Inference*, *148*, 67–81.
Honda, T., Yabe, R. (2017). Variable selection and structure identification for varying coefficient Cox models. *Journal of Multivariate Analysis*, *161*, 103–122.
Huang, J. Z., Wu, C. O., Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, *14*, 763–788.
Ing, C.-K., Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, *22*, 1473–1513.
Javanmard, A., Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, *15*, 2869–2909.
Javanmard, A., Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *Annals of Statistics*, *46*, 2593–2622.
Liu, J., Li, R., Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association*, *109*, 266–274.
Liu, J., Zhong, W., Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, *58*, 1–22.
Lounici, K., van de Pontil, M., Geer, S., Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, *39*, 2164–2204.
Mitra, R., Zhang, C.-H. (2016). The benefit of group sparsity in group inference with de-biased scaled group Lasso. *Electronic Journal of Statistics*, *10*, 1829–1873.
Schumaker, L. L. (2007). *Spline functions*: *Basic theory* 3rd ed. Cambridge: Cambridge University Press.
Stucky, B., van de Geer, S. (2018). Asymptotic confidence regions for high-dimensional structured sparsity. *IEEE Transactions on Signal Processing*, *66*, 2178–2190.
Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.
van de Geer, S. (2016). *Estimation and testing under sparsity*. Dordrecht: Springer.
van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, *42*, 1166–1202.
Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, *104*, 1512–1524.
Wei, F., Huang, J., Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, *21*, 1515–1540.
Yang, Y., Zou, H. (2017). gglasso: Group Lasso penalized learning using a unified BMD algorithm. *R Package Version*, *1*, 4.
Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, *68*, 49–67.
Zhang, C.-H., Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, *76*, 217–242.
Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.