Wang and Wang *Journal of Inequalities and Applications* (2017) 2017:158
DOI 10.1186/s13660-017-1432-x

Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Adaptive group bridge estimation for high-dimensional partially linear models

Xiuli Wang and Mingqiu Wang*

*Correspondence:
wmq0829@gmail.com
School of Statistics, Qufu Normal University, Jingxuan West Road, Qufu, 273165, P.R. China

Abstract

This paper studies group selection for the partially linear model with a diverging number of parameters. We propose an adaptive group bridge method and study the consistency, convergence rate and asymptotic distribution of the global adaptive group bridge estimator under regularity conditions. Simulation studies and a real example show the finite sample performance of our method.

MSC: 62E20; 62J07; 62F12

Keywords: adaptive group bridge; high dimension; partially linear model

1 Introduction

Consider the following model:

$$Y = \mathbf{x}^T \boldsymbol{\beta} + f(U) + \varepsilon, \tag{1}$$

where $\mathbf{x} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_{p_n}^T)^T$ is a covariate vector with $\boldsymbol{x}_j = (X_{jk}, k = 1, \ldots, d_j)^T$ being a $d_j \times 1$ vector corresponding to the jth group in the linear part, $\boldsymbol{\beta} = (\boldsymbol{\beta}_j^T, j = 1, \ldots, p_n)^T$ with $\boldsymbol{\beta}_j$ being the $d_j \times 1$ vector of regression coefficients, f is an unknown function of U, and ε is the random error with mean zero. Without loss of generality, U is scaled to $[0,1]$. Furthermore, $(\mathbf{x}, U)$ and ε are independent.

Variable selection for high-dimensional data is a hot and important issue. Penalized regression methods have been widely used in the literature such as [1–5], and so on. Among these methods, bridge regression including lasso and ridge as two well-known special cases has been studied by many authors (e.g., [6–10]). [11] studied adaptive bridge estimation for high-dimensional linear models. In addition, group structure of variables arise always in many contemporary statistical modeling problems. [12] proposed a group bridge method which not only effectively removes unimportant groups, but also maintains the flexibility of selecting variables within identified groups. [13] investigated an adaptive choice of the penalty order in group bridge regression.

The aforementioned model (1) is just the partially linear model that originated from [14]. The partially linear model is a common semiparametric model enjoying the interpretability and flexibility. Our contributions in this paper include: (1) we propose an adaptive group bridge method to achieve the group selection for a high-dimensional partially linear model; (2) we consider the choice of index γ in the adaptive group bridge and use leave-one-observation-out cross-validation (CV) to implement this choice. It can significantly reduce the computational burden; (3) we give the consistency, convergence rate and asymptotic distribution of the adaptive group bridge estimator which is the global minimizer of the objective function.

The rest of the article is organized as follows. Section 2 gives the adaptive group bridge method. In Section 3, we show the assumptions and asymptotic results for the global adaptive group bridge estimator. Section 4 shows computational algorithm and selection of tuning parameters. Simulation studies and real data are presented in Section 5. Section 6 gives a short discussion. Technical proofs are relegated to Appendix.

2 Adaptive group bridge in the partially linear model

Suppose that we have a collection of independent observations $\{(\mathbf{x}_i, U_i, Y_i), 1 \leq i \leq n\}$ from model (1). That is,

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + f(U_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random errors with mean zero and finite variance $\sigma^2 < \infty$.

To obtain an estimate of function $f(\cdot)$, we employ a B-spline basis. Denote $\mathcal{S}_n$ as the space of polynomial splines of degree $m \geq 1$. Let $\{B_k(u), 1 \leq k \leq q_n\}$ be a normalized B-spline basis with $\|B_k\|_\infty \leq 1$, where $\| \cdot \|_\infty$ is the sup norm. Then, for any $f_n \in \mathcal{S}_n$, we have

$$f_n(u) = \sum_{j=1}^{q_n} B_j(u)\alpha_j \triangleq \mathbf{B}(u)^T \boldsymbol{\alpha}.$$

Under some smoothness conditions, the nonparametric function f can well be approximated by functions in $\mathcal{S}_n$.

Consider the following adaptive group bridge penalized objective function:

$$\sum_{i=1}^{n} \bigl(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \mathbf{B}(U_i)^T \boldsymbol{\alpha}\bigr)^2 + \sum_{j=1}^{p_n} \lambda_j \|\boldsymbol{\beta}_j\|^\gamma, \tag{3}$$

where λ_j, $j = 1, \ldots, p_n$, are the tuning parameters, and $\| \cdot \|$ denotes the L_2 norm on the Euclidean space. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbb{X} = (X_{ijk}, 1 \leq i \leq n, 1 \leq j \leq p_n, 1 \leq k \leq d_j) = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{Z} = (\mathbf{B}(U_1), \ldots, \mathbf{B}(U_n))^T$. Then (3) can be changed into

$$L_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \|\mathbf{Y} - \mathbb{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}\|^2 + \sum_{j=1}^{p_n} \lambda_j \|\boldsymbol{\beta}_j\|^\gamma. \tag{4}$$

For some $\boldsymbol{\beta}$, the optimal $\boldsymbol{\alpha}$ minimizing $L_n(\cdot)$ meets the partial differential equation

$$\partial L_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \boldsymbol{\alpha} = 0,$$

namely,

$$\mathbf{Z}^T \mathbf{Z} \boldsymbol{\alpha} = \mathbf{Z}^T (\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}).$$

Let $H = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, note that H is a projection matrix. We can rewrite the expression (4) as follows:

$$Q_n(\boldsymbol{\beta}) = \left\|(I-H)(\mathbf{Y}-\mathbb{X}\boldsymbol{\beta})\right\|^2 + \sum_{j=1}^{p_n}\lambda_j\|\boldsymbol{\beta}_j\|^{\gamma}. \tag{5}$$

For some fixed $\gamma > 0$, define $\hat{\boldsymbol{\beta}} = \arg\min Q_n(\boldsymbol{\beta})$, then $\hat{\boldsymbol{\beta}}$ is called the adaptive group bridge estimator. If $\hat{\boldsymbol{\beta}}$ is obtained, then the estimator $\hat{\boldsymbol{\alpha}}$ can be achieved. Thus we can get the estimator of the nonparametric part, namely, $\hat{f}_n(u) = \mathbf{B}(u)^T\hat{\boldsymbol{\alpha}}$.

3 Asymptotic properties

In this section, we show the oracle property of the parametric part. For convenience of the statement, we first give some notations. Define $\mathbf{g}(u) = E(\mathbf{x}|U = u)$ and $\tilde{\mathbf{x}} = \mathbf{x} - E(\mathbf{x}|U)$. Let $\Sigma(u)$ be the conditional covariance matrix of $\tilde{\mathbf{x}}$, i.e., $\Sigma(u) = \mathrm{cov}(\tilde{\mathbf{x}}|U = u)$. Denote Ω as the unconditional covariance matrix of $\tilde{\mathbf{x}}$, i.e., $\Omega = E[\Sigma(U)]$. The corresponding sample version is $\mathbf{G} = (\mathbf{g}(U_1),\dots,\mathbf{g}(U_n))^T$ with $\mathbf{g}(U_i) = E(\mathbf{x}_i|U_i)$ and $\widetilde{\mathbb{X}} = (\tilde{\mathbf{x}}_1,\dots,\tilde{\mathbf{x}}_n)^T$ with $\tilde{\mathbf{x}}_i = \mathbf{x}_i - E(\mathbf{x}_i|U_i)$.

Let the true parameter be $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T,\dots,\boldsymbol{\beta}_{0p_n}^T)^T \triangleq (\boldsymbol{\beta}_{10}^T,\boldsymbol{\beta}_{20}^T)^T$. Let $\mathcal{A} = \{1 \le j \le p_n : \|\boldsymbol{\beta}_{0j}\| \neq 0\}$ be the index set of the nonzero groups. Without loss of generality, we assume that coefficients of the first k_n group are nonzero, i.e., $\mathcal{A} = \{1,2,\dots,k_n\}$. Let $|\mathcal{A}| = k_n$ be the cardinality of the set $\mathcal{A}$, which is allowed to increase with n. For $j \notin \mathcal{A}$, $\|\boldsymbol{\beta}_{0j}\| = 0$. Define $\boldsymbol{\beta}_{10} = (\boldsymbol{\beta}_{0j}^T, j \in \mathcal{A})^T$, $\boldsymbol{\beta}_{20} = (\boldsymbol{\beta}_{0j}^T, j \notin \mathcal{A})^T$. Let $d^* = \max_{1\le j\le p_n} d_j$, $\varphi_{n1} = \max\{\lambda_j, j \in \mathcal{A}\}$ and $\varphi_{n2} = \min\{\lambda_j, j \notin \mathcal{A}\}$.

Corresponding to the partition of $\boldsymbol{\beta}_0$, denote $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{(1)}^T, \hat{\boldsymbol{\beta}}_{(2)}^T)^T$ and decompose

$$\mathbb{X} = (\mathbb{X}_1\mathbb{X}_2), \qquad \mathbf{G} = (\mathbf{G}_1\mathbf{G}_2), \qquad \widetilde{\mathbb{X}} = (\widetilde{\mathbb{X}}_1\widetilde{\mathbb{X}}_2), \qquad \Omega = \begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{pmatrix}.$$

The following conditions are required for the B-spline approximation of function f.

(C1) The distribution of U is absolutely continuous, and its density is bounded away from 0 and ∞.

(C2) (Hölder conditions of $f(\cdot)$ and $g_j(\cdot)$, where g_j is the jth component of $\mathbf{g}$) Let l, δ and M be real constants such that $0 < \delta \le 1$ and $M > 0$. $f(\cdot)$ and $g_j(\cdot)$ belong to a class of functions $\mathcal{H}$,

$$\mathcal{H} = \left\{h : \left|h^{(l)}(u_1) - h^{(l)}(u_2)\right| \le M|u_1 - u_2|^{\delta}, \text{for } 0 \le u_1, u_2 \le 1\right\},$$

where $0 < l \le m - 1$ and $r = l + \delta$.

The following part lists all the reasonable conditions which are necessary to attain the asymptotic results.

(A1) Let $\lambda_{\max}(\Omega)$ and $\lambda_{\min}(\Omega)$ be the largest and smallest eigenvalue of Ω, respectively. There exist constants τ_1 and τ_2 such that

$$0 < \tau_1 \le \lambda_{\min}(\Omega) \le \lambda_{\max}(\Omega) \le \tau_2 < \infty.$$

(A2) There exist constants $0 < b_0 < b_1 < \infty$ such that

$$b_0 \le \min\{\|\boldsymbol{\beta}_{0j}\|, 1 \le j \le k_n\} \le \max\{\|\boldsymbol{\beta}_{0j}\|, 1 \le j \le k_n\} \le b_1.$$

(A3) $\|n^{-1}\mathbb{X}^T(I-H)\mathbb{X} - \Omega\| \xrightarrow{P} 0$; $E[\mathrm{tr}(\mathbb{X}^T(I-H)\mathbb{X})] = O(np_n)$.

(A4) $d^* = O(1)$, $p_n^2/n \to 0$ and $n^{-1}\varphi_{n1}k_n \to 0$.

(A5) (a) $\varphi_{n1}k_n^{1/2}/(\sqrt{np_n} + n\sqrt{p_n}q_n^{-r}) \to 0$; (b) $\varphi_{n2}(\sqrt{n^{-1}p_n} + \sqrt{p_n}q_n^{-r})^{\gamma-2}/n \to \infty$.

(A6) For every $1 \le j \le p_n$ and $1 \le k \le d_j$, $E[X_{1jk} - E(X_{1jk}|U_1)]^4$ is bounded. Furthermore, $E(\varepsilon^4)$ is bounded.

Conditions (A1) and (A2) are commonly used. Condition (A3) holds under some conditions. The proof can be found in Lemmas 1 and 2 in [15]. Condition (A4) is used to obtain the consistency of the estimator. Condition (A5) is needed in the proof of convergence rate. Condition (A6) is necessary to attain the asymptotic distribution.

Theorem 3.1 (Consistency) *Suppose that* $\gamma > 0$ *and conditions* (A1)-(A4) *hold, then*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 = O_P\big(n^{-1}d^*p_n + q_n^{-2r} + n^{-1}\varphi_{n1}k_n\big),$$

namely, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

Theorem 3.1 implies that under some conditions the estimators converge to the true values of parameters.

Theorem 3.2 (Convergence rate) *Suppose that conditions* (A1)-(A5) *hold, then*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P\big(\sqrt{n^{-1}p_n} + \sqrt{p_n}q_n^{-r}\big).$$

This theorem shows that the adaptive group bridge can give the optimal convergence rate with $p_n \to \infty$.

Theorem 3.3 (Oracle property) *Suppose that* $0 < \gamma < 1$, $n^{-1}k_nq_n \to 0$ *and* $nq_n^{-2r} \to 0$. *If conditions* (A1)-(A6) *are satisfied, then we have*

(i) $\Pr(\hat{\boldsymbol{\beta}}_{(2)} = \mathbf{0}) \to 1$, $n \to \infty$;

(ii) *Let* $u_n^2 = n^2\boldsymbol{\omega}_n^T(\mathbb{X}_1^T(I-H)\mathbb{X}_1)^{-1}\Omega_{11}(\mathbb{X}_1^T(I-H)\mathbb{X}_1)^{-1}\boldsymbol{\omega}_n$ *with* $\boldsymbol{\omega}_n$ *being some* $\sum_{j=1}^{k_n} d_j$*-vector with* $\|\boldsymbol{\omega}_n\|^2 = 1$, *then*

$$n^{1/2}u_n^{-1}\boldsymbol{\omega}_n^T(\hat{\boldsymbol{\beta}}_{(1)} - \boldsymbol{\beta}_{10}) \xrightarrow{D} N\big(0, \sigma^2\big).$$

This theorem states that the adaptive group bridge performs as well as the oracle [16].

4 Computational algorithm and selection of tuning parameters

4.1 Computational algorithm

In this section, we apply the LQA algorithm proposed by [3] to compute the adaptive group bridge estimate.

We take the initial value $\boldsymbol{\beta}^{(0)}$. Here the ordinary least square estimate is chosen as the initial value $\boldsymbol{\beta}^{(0)}$. The penalty term $p_{\lambda_j}(\|\boldsymbol{\beta}_j\|) = \lambda_j \|\boldsymbol{\beta}_j\|^{\gamma}$ can be approximated as

$$p_{\lambda_j}(\|\boldsymbol{\beta}_j\|) \approx p_{\lambda_j}(\|\boldsymbol{\beta}_j^{(0)}\|) + \frac{1}{2}\{p'_{\lambda_j}(\|\boldsymbol{\beta}_j^{(0)}\|)/\|\boldsymbol{\beta}_j^{(0)}\|\}(\|\boldsymbol{\beta}_j\|^2 - \|\boldsymbol{\beta}_j^{(0)}\|^2),$$

when $\|\boldsymbol{\beta}_j^{(0)}\| > 0$. The following iterative expression of $\boldsymbol{\beta}$ can be obtained:

$$\boldsymbol{\beta}^{(1)} = \left[\mathbb{X}^T(I-H)\mathbb{X} + n\Sigma_{\lambda,\gamma}(\boldsymbol{\beta}^{(0)})\right]^{-1}\mathbb{X}^T(I-H)\mathbf{Y}, \tag{6}$$

where

$$\Sigma_{\lambda,\gamma}(\boldsymbol{\beta}^{(0)}) = \operatorname{diag}\left\{\frac{p'_{\lambda_j}(\|\boldsymbol{\beta}_j^{(0)}\|)}{\|\boldsymbol{\beta}_j^{(0)}\|} I_{d_j}, j = 1, \ldots, p_n\right\},$$

with I_{d_j} being a $d_j \times d_j$ unit matrix. If some $\|\boldsymbol{\beta}_j^{(1)}\|$ is smaller than 10^{-3}, then we set $\boldsymbol{\beta}_j^{(1)} = \mathbf{0}$. The finial estimate can be obtained iteratively by formula (6) until the convergence is achieved.

4.2 Selection of the tuning parameters

For our method, q_n, γ, and λ_j ($j = 1, \ldots, p_n$) should be chosen. For convenience, cubic spline basis ($m = 4$) is used. We set $q_n = 7$. Simulation results demonstrate that this choice performs quite well. There are also many tuning parameters that should be chosen. In fact, we only need to select one tuning parameter by setting $\lambda_j = \lambda/\|\boldsymbol{\beta}_j^{(0)}\|$. We use 'leave-one-observation-out' cross-validation (CV) to select λ and γ. Due to the convergence of the algorithm, we have

$$\hat{\boldsymbol{\beta}} = \left[\mathbb{X}^T(I-H)\mathbb{X} + n\Sigma_{\lambda,\gamma}(\hat{\boldsymbol{\beta}})\right]^{-1}\mathbb{X}^T(I-H)\mathbf{Y},$$

where $\hat{\boldsymbol{\beta}}$ is obtained based on the whole data set. Note that it is the solution of the ridge regression

$$\left\|\mathbf{Y}^* - \mathbb{X}^*\boldsymbol{\beta}\right\|^2 + n\boldsymbol{\beta}^T\Sigma_{\lambda,\gamma}(\hat{\boldsymbol{\beta}})\boldsymbol{\beta}, \tag{7}$$

where $\mathbf{Y}^* = (I-H)\mathbf{Y}$ and $\mathbb{X}^* = (I-H)\mathbb{X}$. Let $\mathbf{Y}^* = (y_1^*, \ldots, y_n^*)^T$ and $\mathbb{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)^T$. The CV error is

$$CV(\lambda,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - \mathbf{x}_i^{*T}\hat{\boldsymbol{\beta}}^{-i}\right)^2,$$

where $\hat{\boldsymbol{\beta}}^{-i}$ is achieved by solving (7) without the ith observation. The computation of the CV error is intensive, so we will use the following formula, which can be proved similar to [17]:

$$CV(\lambda,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^* - \mathbf{x}_i^{*T}\hat{\boldsymbol{\beta}}\right)^2/(1-D_{ii}),$$

where D_{ii} is the (i,i)th diagonal element of $(I-H)\mathbb{X}[\mathbb{X}^T(I-H)\mathbb{X}+n\Sigma_{\lambda,\gamma}(\hat{\boldsymbol{\beta}})]^{-1}\mathbb{X}^T(I-H)$. It is obvious that this method can significantly reduce the computational burden.

5 Simulation studies and application

In this section, we investigate the finite sample performance of the adaptive group bridge method through simulations and a real data application.

5.1 Monte Carlo simulations

We simulate 100 datasets consisting of n observations from the following partially linear model:

$$Y_i=\sum_{j=1}^{p_n}\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j+\cos(2\pi U_i)+\varepsilon_i,\quad i=1,\ldots,n,$$

where $n=500$, and the error $\varepsilon_i\sim N(0,\sigma^2)$ with $\sigma=0.5,1,4$. We consider that there are p_n groups with $p_n=10,30,50$ and each group consists of three variables. The true values of parameters $\boldsymbol{\beta}_1^T=(0.5,1,1.5)$, $\boldsymbol{\beta}_2^T=(1,-1,1)$, $\boldsymbol{\beta}_3^T=(0.5,0.5,0.5)$, $\boldsymbol{\beta}_4^T=\cdots=\boldsymbol{\beta}_{p_n}^T=(0,0,0)$. U_i follows the uniform distribution on $[0,1]$. To generate covariate $\mathbf{x}=(\boldsymbol{x}_1^T,\boldsymbol{x}_2^T,\ldots,\boldsymbol{x}_{p_n}^T)^T$ with $\boldsymbol{x}_j=(X_{jk},k=1,2,3)^T$, we first simulate $R_1,\ldots,R_{3p_n}$ independently from the standard normal distribution. Next, simulate Z_j, $j=1,\ldots,p_n$, from a multivariate normal distribution with the mean zero and $\mathrm{Cov}(Z_j,Z_l)=0.6^{|j-l|}$. Then the covariates are generated as $X_{jk}=(Z_j+R_{3(j-1)+k})/\sqrt{2}$, $j=1,\ldots,p_n$, $k=1,2,3$.

We compare the adaptive group bridge (AGB) with the group lasso (GL) and the group bridge (GB). The following three performance measures are calculated:

1. L_2 loss of parametric estimate, which is defined as $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|$.
2. Average number of nonzero groups identified by the method (NN).
3. Average number of nonzero groups identified by the method that are truly nonzero (NNT).

Group selection results are depicted in Table 1. The numbers in the parentheses in the columns labeled 'NN' and 'NNT' are the corresponding sample standard deviations based on the 100 runs. Boxplots of the L_2 losses under different settings are given in Figures 1-3.

From Table 1, we can have the following observations:

(1) Both GB and AGB perform better than GL for all settings. All these three methods can retain all the true nonzero groups, but GL always keeps more redundant groups that are unrelated with the response than both GB and AGB.

(2) AGB performs much better for larger σ and p_n. When $p_n=50$ for AGB, groups selected for the case $\sigma=4$ are about 18.5% lower than that for the case $\sigma=0.5$. While groups selected for GB decrease by 7.37% in the same situation.

(3) For $p_n=10$, GB performs better than AGB, but the stability of GB is bad for $\sigma=4$.

Figures 1-3 present L_2 losses with varying σ and p_n. We can see that the performances of estimates are similar for GB and AGB. For $p_n=30$ and 50, both GB and AGB perform better than GL. However, when $p_n=50$, the median of L_2 losses for all these three are similar for $\sigma=0.5$ and 4, but the L_2 losses of GL fluctuate more widely.

Table 1 Group selection results

p_n	Method	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 4$	
		NN	NNT	NN	NNT	NN	NNT
10	GL	7.80 (1.231)	3 (0)	7.59 (2.396)	3 (0)	6.02 (1.255)	3 (0)
	GB	4.64 (1.259)	3 (0)	4.80 (1.356)	3 (0)	4.55 (1.258)	3 (0)
	AGB	5.22 (1.605)	3 (0)	5.00 (1.735)	3 (0)	4.56 (1.131)	3 (0)
30	GL	20.47 (2.504)	3 (0)	11.49 (4.464)	3 (0)	14.46 (2.022)	3 (0)
	GB	11.04 (3.643)	3 (0)	10.96 (3.784)	3 (0)	10.18 (2.350)	3 (0)
	AGB	13.06 (5.510)	3 (0)	10.64 (5.921)	3 (0)	9.57 (2.046)	3 (0)
50	GL	33.17 (3.223)	3 (0)	16.48 (5.018)	3 (0)	22.37 (3.308)	3 (0)
	GB	17.23 (4.608)	3 (0)	17.79 (3.952)	3 (0)	15.96 (3.284)	3 (0)
	AGB	19.25 (8.437)	3 (0)	15.67 (8.385)	3 (0)	15.68 (3.684)	3 (0)

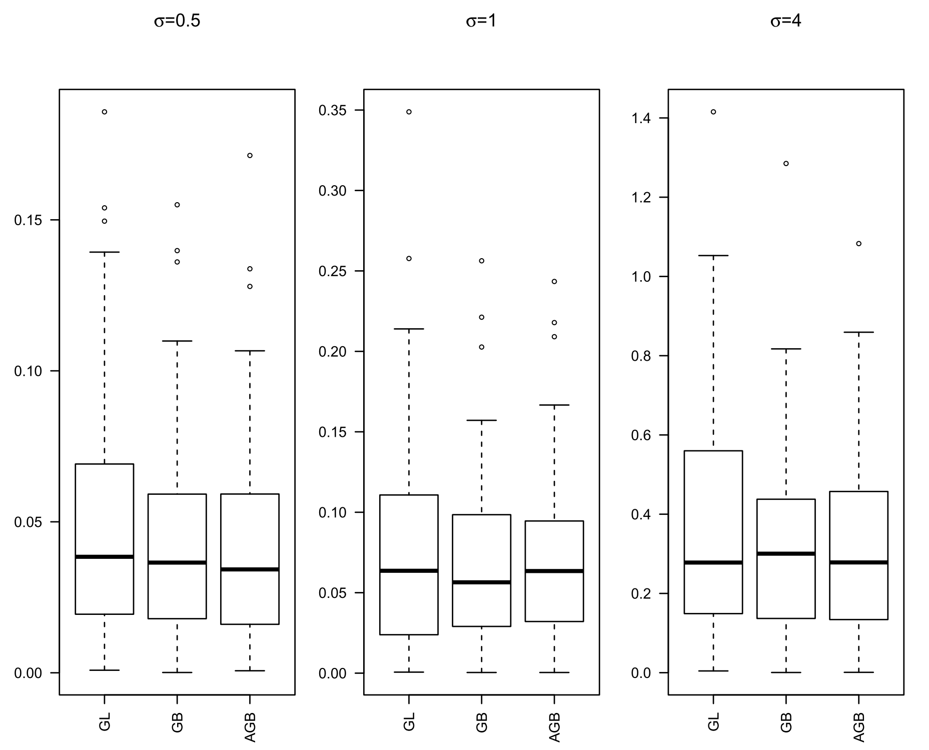

Figure 1 Boxplots of L_2 loss for $p_n = 10$.

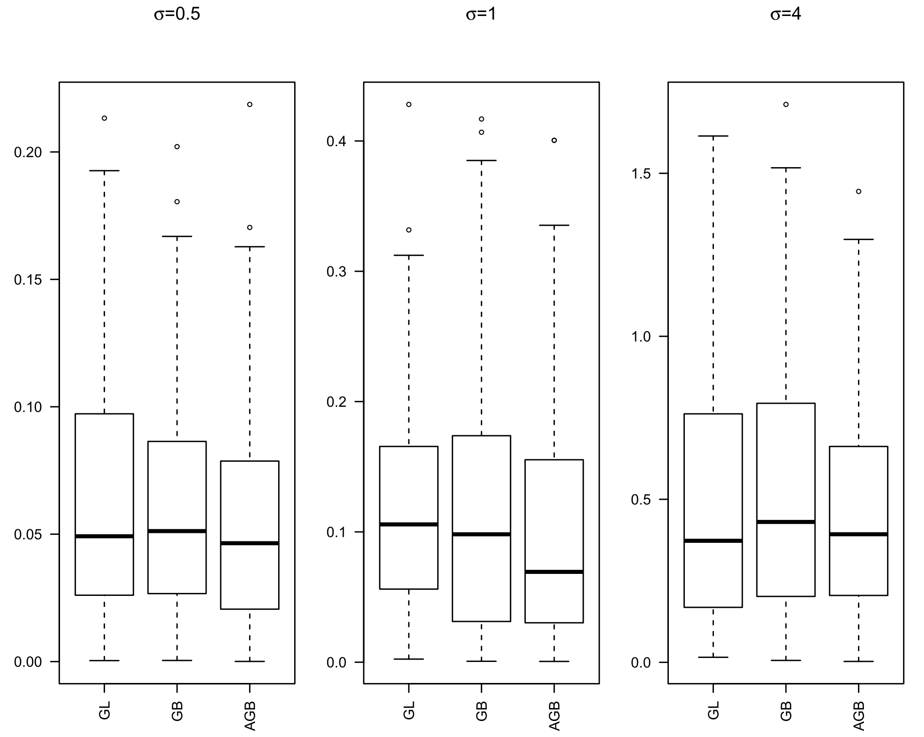

Figure 2 Boxplots of L_2 loss for $p_n = 30$.

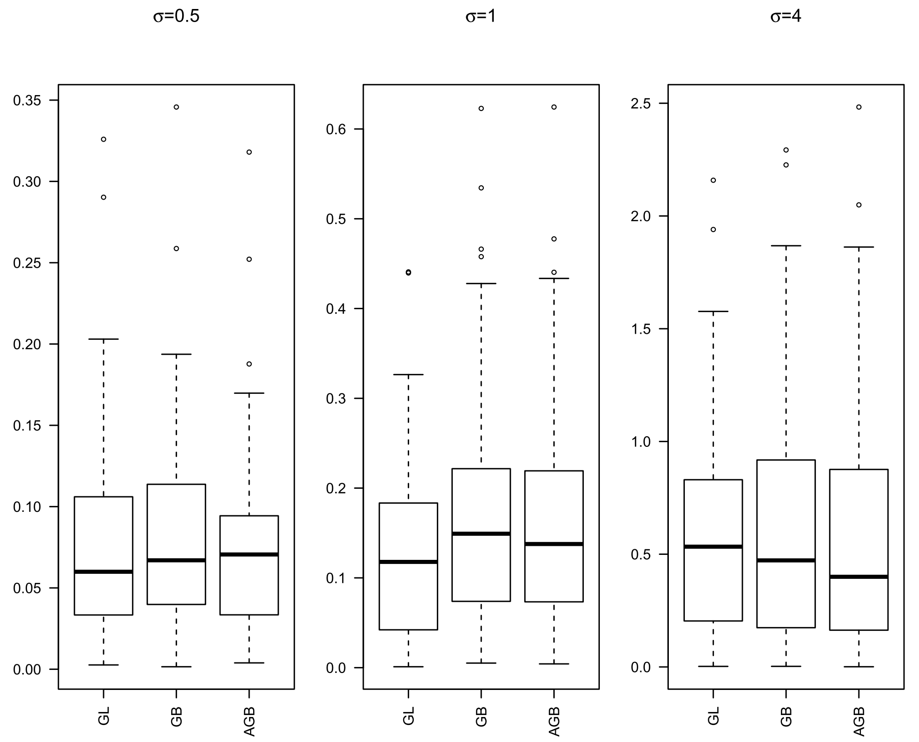

Figure 3 Boxplots of L_2 loss for $p_n = 50$.

Table 2 Estimates of the wage data

Variable	Description	GL	GB	AGB
edu	Number of years of education	0.0694	0.0668	0.0635
south	1 = southern region, 0 = other	−0.0723	−0.0679	−0.0490
sex	1 = Female, 0 = Male	−0.1999	−0.1983	−0.2031
union	1 = union member, 0 = nonmember	0.1951	0.1934	0.2030
race	1 = other, 0 = White	−0.0559	−0.0585	−0.0582
	1 = Hispanic, 0 = White	−0.0537	−0.0615	−0.0614
occup	1 = management, 0 = other	0.1874	0.2173	0.2516
	1 = sales, 0 = other	−0.0797	−0.0809	−0.0721
	1 = clerical, 0 = other	0.0166	0.0262	0.0430
	1 = service, 0 = other	−0.1171	−0.1173	−0.1104
	1 = professional, 0 = other	0.1533	0.1768	0.2061
sector	1 = manufacturing, 0 = other	0.0848	0.0912	0.0994
	1 = construction, 0 = other	0.0546	0.0622	0.0674
marr	1 = married, 0 = other	0.0000	0.0000	0.0000

5.2 Wage data analysis

The workers' wage data from Berndt[18] contains a random sample of 534 observations on 11 variables sampled from the current population survey of 1985. It provides information on wages and other characteristics of the workers, including continuous variables: the number of years of education, years of work experience, age and nominal variables: race, sex, region of residence, occupational status, sector, marital status and union membership. Our goal is to study the important factors for the wage, so it is reasonable to use our proposed method for these data.

From the residual plot, we can easily see that the variance of wages is not a constant. So the *log* transformation is used to stabilize the variance of wages. Due to the multicollinearity problem between age and experience, we need to get rid of either age or experience. Here we remove the age variable from the model. Xie and Huang [15] analyzed these data without considering the transformation of Y. Furthermore, they did not consider group selection of factors. Similar to Xie and Huang [15], we fit these data using a partially linear model with U being 'years of work experience'.

Table 2 reports estimated regression coefficients of GL, GB and AGB. All these three methods exclude marital status. We use the first 400 observations as a training dataset to select and fit the model, and use the rest of 134 observations as a testing dataset to evaluate the prediction ability of the selected model. The prediction performance is measured by the median of $\{|y_i - \hat{y}_i|, i = 1, 2, \dots, 134\}$ for GL, GB and AGB using the testing data, respectively. Here y_i's are those 134 observations in the testing dataset and $\hat{y}_i$'s are corresponding prediction values. The median absolute prediction errors of GL, GB and AGB are 0.3072, 0.3062 and 0.3022, respectively. Therefore, we can conclude that the AGB gives the smallest prediction error, so it is an attractive technique in group selection.

6 Discussion

This paper studies group selection for high-dimensional partially linear model with the adaptive group bridge method. We also consider the choice of γ in the bridge penalty. It is worth mentioning that we use 'leave-one-observation-out' cross-validation to select both λ and γ. This method can significantly reduce the computational burden. This is the first try to use this method in group selection for the partially linear model.

Appendix

Proof of Theorem 3.1 By the definition of $\hat{\boldsymbol{\beta}}$, it is easy to get

$$\left\|(I-H)(\mathbf{Y}-\mathbb{X}\hat{\boldsymbol{\beta}})\right\|^2+\sum_{j=1}^{p_n}\lambda_j\|\hat{\boldsymbol{\beta}}_j\|^\gamma\leq\left\|(I-H)(\mathbf{Y}-\mathbb{X}\boldsymbol{\beta}_0)\right\|^2+\sum_{j=1}^{p_n}\lambda_j\|\boldsymbol{\beta}_{0j}\|^\gamma,$$

that is,

$$\left\|(I-H)(\mathbf{Y}-\mathbb{X}\hat{\boldsymbol{\beta}})\right\|^2-\left\|(I-H)(\mathbf{Y}-\mathbb{X}\boldsymbol{\beta}_0)\right\|^2\leq\sum_{j=1}^{p_n}\lambda_j\|\boldsymbol{\beta}_{0j}\|^\gamma.$$

As $\mathbf{Y}=\mathbb{X}\boldsymbol{\beta}_0+\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}$ with $\boldsymbol{f}(\mathbf{U})=(f(U_1),\ldots,f(U_n))^T$ and $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^T$, we can rewrite the upper inequality as follows:

$$\left\|(I-H)\mathbb{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right\|^2-2\big(\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}\big)^T(I-H)\mathbb{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\leq\sum_{j=1}^{p_n}\lambda_j\|\boldsymbol{\beta}_{0j}\|^\gamma.$$

Let

$$\begin{aligned}a_n&=n^{-1/2}\big[\mathbb{X}^T(I-H)\mathbb{X}\big]^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0),\\ b_n&=n^{-1/2}\big[\mathbb{X}^T(I-H)\mathbb{X}\big]^{-1/2}\mathbb{X}^T(I-H)\big(\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}\big).\end{aligned}$$

Then we have

$$\|a_n\|^2\leq2\big(\|a_n-b_n\|^2+\|b_n\|^2\big)\leq\frac{2}{n}\sum_{j=1}^{p_n}\lambda_j\|\boldsymbol{\beta}_{0j}\|^\gamma+4\|b_n\|^2.$$

Since $|\mathcal{A}|=k_n$, under condition (A2),

$$\frac{2}{n}\sum_{j=1}^{p_n}\lambda_j\|\boldsymbol{\beta}_{0j}\|^\gamma=O\left(\frac{\varphi_{n1}k_n}{n}\right).$$

While

$$\begin{aligned}\|b_n\|^2&=\frac{1}{n}\big(\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}\big)^T(I-H)\mathbb{X}\big[\mathbb{X}^T(I-H)\mathbb{X}\big]^{-1}\mathbb{X}^T(I-H)\big(\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}\big)\\&\leq\frac{2}{n}\boldsymbol{\varepsilon}^TA\boldsymbol{\varepsilon}+\frac{2}{n}\boldsymbol{f}(\mathbf{U})^TA\boldsymbol{f}(\mathbf{U}),\end{aligned}\tag{8}$$

where

$$A=(I-H)\mathbb{X}\big[\mathbb{X}^T(I-H)\mathbb{X}\big]^{-1}\mathbb{X}^T(I-H).$$

For the first term on the right-hand side of (8),

$$E\left(\frac{1}{n}\boldsymbol{\varepsilon}^TA\boldsymbol{\varepsilon}\right)=\frac{\sigma^2}{n}\operatorname{tr}\big(E(A)\big)\leq n^{-1}d^*p_n\sigma^2.$$

Thus

$$n^{-1}\boldsymbol{\varepsilon}^T A\boldsymbol{\varepsilon} = O_P\bigl(n^{-1}d^* p_n\bigr). \tag{9}$$

For the second term on the right-hand side of (8), by conditions (C1) and (C2),

$$\begin{aligned} E\biggl(\frac{1}{n}\boldsymbol{f}(\mathbf{U})^T A\boldsymbol{f}(\mathbf{U})\biggr) &\le \frac{1}{n}E\bigl\{\lambda_{\max}\bigl\{(I-H)\mathbb{X}\bigl[\mathbb{X}^T(I-H)\mathbb{X}\bigr]^{-1}\mathbb{X}^T(I-H)\bigr\} \\ &\quad \times \operatorname{tr}\bigl[\boldsymbol{f}(\mathbf{U})^T(I-H)\boldsymbol{f}(\mathbf{U})\bigr]\bigr\} \\ &= \frac{1}{n}E\bigl[\boldsymbol{f}(\mathbf{U})^T(I-H)\boldsymbol{f}(\mathbf{U})\bigr] = O\bigl(q_n^{-2r}\bigr). \end{aligned} \tag{10}$$

Combining (9)-(10),

$$\|b_n\|^2 = O_P\bigl(n^{-1}d^* p_n + q_n^{-2r}\bigr).$$

By conditions (A1) and (A3),

$$\begin{aligned} E\|a_n\|^2 &= \frac{1}{n}E\bigl[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\mathbb{X}^T(I-H)\mathbb{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\bigr] \\ &= E\biggl[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\biggl(\frac{1}{n}\mathbb{X}^T(I-H)\mathbb{X}-\Omega\biggr)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\biggr] \\ &\quad + E\bigl[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\Omega(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\bigr] \\ &\ge \frac{\tau_1}{2}E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2. \end{aligned}$$

Therefore

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2 = O_P\bigl(n^{-1}d^* p_n + q_n^{-2r} + n^{-1}\varphi_{n1}k_n\bigr).$$

Under condition (A4), we have

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| \xrightarrow{P} 0. \qquad \square$$

Proof of Theorem 3.2 Let $\mu_n = \sqrt{n^{-1}p_n} + q_n^{-r} + \sqrt{n^{-1}\varphi_{n1}k_n}$, we can choose a sequence $\{r_n, r_n > 0\}$ which satisfies $r_n \to 0$. Partition $\mathbb{R}^{\sum_{j=1}^{p_n} d_j}\backslash\{0\}$ into shells $\{S_{nj} : j = 1, 2, \ldots\}$, where $S_{nj} = \{\boldsymbol{\beta} : 2^{j-1}r_n \le \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| < 2^j r_n\}$. For an arbitrary fixed constant $L \in \mathbb{R}^+$, if $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|$ is larger than $2^L r_n$, $\hat{\boldsymbol{\beta}}$ is in one of the shells with $j \ge L$, we have

$$\begin{aligned} \Pr\bigl(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| \ge 2^L r_n\bigr) &= \sum_{l>L, 2^l r_n > 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) \\ &\quad + \sum_{l>L, 2^l r_n \le 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) \quad (L_1 \text{ is an arbitrary constant}), \end{aligned}$$

where

$$\sum_{l>L, 2^l r_n > 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) \le \Pr\bigl(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| \ge 2^{L_1-1}\mu_n\bigr) = o(1),$$

and

$$\begin{aligned}
&\sum_{l>L,2^l r_n\le 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol\beta}\in S_{nl})\\
&\quad=\sum_{l>L,2^l r_n\le 2^{L_1}\mu_n} \Pr\biggl(\hat{\boldsymbol\beta}\in S_{nl}, \|\Delta_n\|\le\frac{\tau_1}{2}\biggr)\\
&\qquad+\sum_{l>L,2^l r_n\le 2^{L_1}\mu_n} \Pr\biggl(\hat{\boldsymbol\beta}\in S_{nl}, \|\Delta_n\|>\frac{\tau_1}{2}\biggr),
\end{aligned}$$

where $\Delta_n = n^{-1}\mathbb{X}^T(I-H)\mathbb{X}-\Omega$. By condition (A3),

$$\sum_{l>L,2^l r_n\le 2^{L_1}\mu_n} \Pr\biggl(\hat{\boldsymbol\beta}\in S_{nl}, \|\Delta_n\|>\frac{\tau_1}{2}\biggr)\le \Pr\biggl(\|\Delta_n\|>\frac{\tau_1}{2}\biggr)=o(1).$$

Therefore,

$$\begin{aligned}
&\Pr\bigl(\|\hat{\boldsymbol\beta}-\boldsymbol\beta_0\|\ge 2^L r_n\bigr)\\
&\quad=o(1)+\sum_{l>L,2^l r_n\le 2^{L_1}\mu_n} \Pr\biggl(\inf_{\boldsymbol\beta\in S_{nl}}\bigl(Q_n(\boldsymbol\beta)-Q_n(\boldsymbol\beta_0)\bigr)<0, \|\Delta_n\|\le\frac{\tau_1}{2}\biggr).
\end{aligned}$$

Since

$$\begin{aligned}
&Q_n(\boldsymbol\beta)-Q_n(\boldsymbol\beta_0)\\
&\quad=\bigl\|(I-H)\mathbb{X}(\boldsymbol\beta-\boldsymbol\beta_0)\bigr\|^2-2\bigl(\boldsymbol f(\mathrm{U})+\boldsymbol\varepsilon\bigr)^T(I-H)\mathbb{X}(\boldsymbol\beta-\boldsymbol\beta_0)\\
&\qquad+\sum_{j=1}^{p_n}\lambda_j\bigl(\|\boldsymbol\beta_j\|^\gamma-\|\boldsymbol\beta_{0j}\|^\gamma\bigr)\\
&\quad\ge\bigl\|(I-H)\mathbb{X}(\boldsymbol\beta-\boldsymbol\beta_0)\bigr\|^2-2\bigl(\boldsymbol f(\mathrm{U})+\boldsymbol\varepsilon\bigr)^T(I-H)\mathbb{X}(\boldsymbol\beta-\boldsymbol\beta_0)\\
&\qquad+\sum_{j=1}^{k_n}\lambda_j\bigl(\|\boldsymbol\beta_j\|^\gamma-\|\boldsymbol\beta_{0j}\|^\gamma\bigr)\\
&\quad\stackrel{\triangle}{=}\mathrm{I}_{n1}+\mathrm{I}_{n2}+\mathrm{I}_{n3}.
\end{aligned}$$

For I_{n1},

$$\mathrm{I}_{n1}\ge\inf_{\boldsymbol\beta\in S_{nl}}\frac{n\tau_1}{2}\|\boldsymbol\beta-\boldsymbol\beta_0\|^2,$$

for all $\boldsymbol\beta\in S_{nl}$, there exists $\|\boldsymbol\beta-\boldsymbol\beta_0\|^2\ge 2^{2l-2}r_n^2$, therefore $\mathrm{I}_{n1}\ge n\tau_1 2^{2l-3}r_n^2$.

For I_{n3}, we have

$$\begin{aligned}
|\mathrm{I}_{n3}|&=\sum_{j=1}^{k_n}\lambda_j\gamma\bigl\|\boldsymbol\beta_j^*\bigr\|^{\gamma-1}\bigl(\|\boldsymbol\beta_j\|-\|\boldsymbol\beta_{0j}\|\bigr)\\
&\le\varphi_{n1}\gamma\sum_{j=1}^{k_n}\bigl\|\boldsymbol\beta_j^*\bigr\|^{\gamma-1}\|\boldsymbol\beta_j-\boldsymbol\beta_{0j}\|,
\end{aligned}$$

where $\boldsymbol{\beta}_j^*$ is between $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_{0j}$. By condition (A2) and since we only need to consider $\boldsymbol{\beta}$ with $\boldsymbol{\beta} \in S_{nl}$, $2^l r_n \le 2^{L_1}\mu_n$, there exists a constant $C_3 > 0$ such that

$$|\mathrm{I}_{n3}| \le C_3 \varphi_{n1}\gamma \sum_{j=1}^{k_n} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{0j}\| \le C_3\varphi_{n1}k_n^{1/2}\gamma\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|.$$

So for all $\boldsymbol{\beta} \in S_{nl}$ such that $|\mathrm{I}_{n3}| \le C_3\varphi_{n1}k_n^{1/2}\gamma 2^l r_n$, by the Markov inequality, we have

$$\begin{aligned}
&\Pr\Bigl(\inf_{\boldsymbol{\beta}\in S_{nl}}\bigl(Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}_0)\bigr) \le 0\Bigr) \\
&\quad\le \Pr\Bigl(\sup_{\boldsymbol{\beta}\in S_{nl}} |\mathrm{I}_{n2}| \ge n\tau_1 2^{2l-3} r_n^2 - C_3\varphi_{n1}k_n^{1/2}\gamma 2^l r_n\Bigr) \\
&\quad\le \frac{E(\sup_{\boldsymbol{\beta}\in S_{nl}}|\mathrm{I}_{n2}|)}{n\tau_1 2^{2l-3}r_n^2 - C_3\varphi_{n1}k_n^{1/2}\gamma 2^l r_n}.
\end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
E\Bigl(\sup_{\boldsymbol{\beta}\in S_{nl}}|\mathrm{I}_{n2}|\Bigr) &\le 2\bigl[E\bigl(\bigl(\boldsymbol{f}(\mathbf{U}) + \boldsymbol{\varepsilon}\bigr)^T(I - H)\mathbb{X}\mathbb{X}^T(I - H)\bigl(\boldsymbol{f}(\mathbf{U}) + \boldsymbol{\varepsilon}\bigr)\bigr)\bigr]^{1/2} \\
&\qquad \times \Bigl[E\Bigl(\sup_{\boldsymbol{\beta}\in S_{nl}}\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2\Bigr)\Bigr]^{1/2} \\
&\le 2^{l+3/2}r_n\bigl[E\bigl(\boldsymbol{\varepsilon}^T(I - H)\mathbb{X}\mathbb{X}^T(I - H)\boldsymbol{\varepsilon}\bigr) \\
&\qquad + E\bigl(\boldsymbol{f}(\mathbf{U})^T(I - H)\mathbb{X}\mathbb{X}^T(I - H)\boldsymbol{f}(\mathbf{U})\bigr)\bigr]^{1/2},
\end{aligned}$$

where

$$E\bigl(\boldsymbol{\varepsilon}^T(I - H)\mathbb{X}\mathbb{X}^T(I - H)\boldsymbol{\varepsilon}\bigr) = \sigma^2 E\bigl(\mathrm{tr}\bigl((I - H)\mathbb{X}\mathbb{X}^T(I - H)\bigr)\bigr) = O(np_n)$$

and

$$\begin{aligned}
&E\bigl[\boldsymbol{f}(\mathbf{U})^T(I - H)\mathbb{X}\mathbb{X}^T(I - H)\boldsymbol{f}(\mathbf{U})\bigr] \\
&\quad\le E\bigl[\mathrm{tr}\bigl(\mathbb{X}^T(I - H)\mathbb{X}\bigr)\,\mathrm{tr}\bigl(\boldsymbol{f}(\mathbf{U})^T(I - H)\boldsymbol{f}(\mathbf{U})\bigr)\bigr] \\
&\quad= O(np_n)O\bigl(nq_n^{-2r}\bigr) = O\bigl(n^2p_nq_n^{-2r}\bigr).
\end{aligned}$$

Accordingly,

$$E\Bigl(\sup_{\boldsymbol{\beta}\in S_{nl}}|\mathrm{I}_{n2}|\Bigr) \le C_4 2^l r_n\bigl(\sqrt{np_n} + n\sqrt{p_n}q_n^{-r}\bigr).$$

Then we can get

$$\sum_{l>L, 2^l r_n \le 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) \le \sum_{l>L}\frac{C_4 2^l r_n(\sqrt{np_n} + n\sqrt{p_n}q_n^{-r})}{n\tau_1 2^{2l-3}r_n^2 - C_3\varphi_{n1}k_n^{1/2}\gamma 2^l r_n}.$$

We choose $r_n = (\sqrt{p_n/n} + \sqrt{p_n}q_n^{-r})$, we have

$$\sum_{l>L,2^l r_n \le 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) = \sum_{l>L} \frac{C_4}{\tau_1 2^{l-3} - C_3\varphi_{n1}k_n^{1/2}\gamma/(\sqrt{np_n} + n\sqrt{p_n}q_n^{-r})}.$$

By condition (A5)(a) $\varphi_{n1}k_n^{1/2}/(\sqrt{np_n} + n\sqrt{p_n}q_n^{-r}) \to 0$, for sufficiently large n,

$$2^{l-3} - C_3\tau_1^{-1}\lambda_j k_n^{1/2}/\big(\sqrt{np_n} + n\sqrt{p_n}M_n^{-rg}\big) \ge 2^{l-4}.$$

Thus

$$\sum_{l>L,2^l r_n \le 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) \le \sum_{l>L} \frac{C_4}{2^{l-4}} \le C_4 2^{-(L-5)}.$$

Let $L \to \infty$, then

$$\sum_{l>L,2^l r_n \le 2^{L_1}\mu_n} \Pr(\hat{\boldsymbol{\beta}} \in S_{nl}) \to 0.$$

Hence

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P\big(\sqrt{n^{-1}p_n} + \sqrt{p_n}q_n^{-r}\big). \qquad \square$$

Proof of Theorem 3.3 (i) By Theorem 3.2, for sufficiently large C_5, $\hat{\boldsymbol{\beta}}$ lies in the ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \nu_n C_5\}$ with probability converging to 1, where $\nu_n = \sqrt{n^{-1}p_n} + \sqrt{p_n}q_n^{-r}$. Let $\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{10} + \nu_n\boldsymbol{v}_1$ and $\boldsymbol{\beta}_{(2)} = \boldsymbol{\beta}_{20} + \nu_n\boldsymbol{v}_2 = \nu_n\boldsymbol{v}_2$ with $\|\boldsymbol{v}\|^2 = \|\boldsymbol{v}_1\|^2 + \|\boldsymbol{v}_2\|^2 \le C_5^2$. Let

$$V_n(\boldsymbol{v}_1, \boldsymbol{v}_2) = Q_n(\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}) - Q_n(\boldsymbol{\beta}_{10}, \mathbf{0}) = Q_n(\boldsymbol{\beta}_{10} + \nu_n\boldsymbol{v}_1, \nu_n\boldsymbol{v}_2) - Q_n(\boldsymbol{\beta}_{10}, \mathbf{0}).$$

Then $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ can be attained by minimizing $V_n(\boldsymbol{v}_1, \boldsymbol{v}_2)$ over $\|\boldsymbol{v}\| \le C_5$, except on an event with probability converging to zero. We only need to show that, for some $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ with $\|\boldsymbol{v}\| \le C_5$, if $\|\boldsymbol{v}_2\| > 0$,

$$\Pr\big(V_n(\boldsymbol{v}_1, \boldsymbol{v}_2) - V_n(\boldsymbol{v}_1, \mathbf{0}) > 0\big) \to 1, \quad n \to \infty.$$

Some simple calculations show that

$$\begin{aligned} V_n(\boldsymbol{v}_1, \boldsymbol{v}_2) - V_n(\boldsymbol{v}_1, \mathbf{0}) &= \nu_n^2\big\|(I-H)\mathbb{X}_2\boldsymbol{v}_2\big\|^2 + 2\nu_n^2(\mathbb{X}_1\boldsymbol{v}_1)^T(I-H)(\mathbb{X}_2\boldsymbol{v}_2) \\ &\quad - 2\nu_n\big(\boldsymbol{f}(\mathbf{U}) + \boldsymbol{\varepsilon}\big)^T(I-H)(\mathbb{X}_2\boldsymbol{v}_2) + \sum_{j\notin\mathcal{A}} \lambda_j\|\nu_n\boldsymbol{v}_{2j}\|^\gamma \\ &\overset{\Delta}{=} \mathrm{II}_{n1} + \mathrm{II}_{n2} + \mathrm{II}_{n3} + \mathrm{II}_{n4}. \end{aligned}$$

For the first two terms II_{n1} and II_{n2},

$$\mathrm{II}_{n1} + \mathrm{II}_{n2} \ge -\nu_n^2\big\|(I-H)\mathbb{X}_1\boldsymbol{v}_1\big\|^2 = -n\nu_n^2 C_5^2\big(o_P(1) + \tau_2\big).$$

For II_{n3}, we have

$$\begin{aligned}
&E\bigl[\bigl(\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}\bigr)^T(I-H)\mathbb{X}_2\boldsymbol{v}_2\bigr]^2\\
&\quad\le 2\bigl\{E\bigl[\boldsymbol{f}(\mathbf{U})^T(I-H)\mathbb{X}_2\boldsymbol{v}_2\boldsymbol{v}_2^T\mathbb{X}_2^T(I-H)\boldsymbol{f}(\mathbf{U})\\
&\qquad+\boldsymbol{\varepsilon}^T(I-H)\mathbb{X}_2\boldsymbol{v}_2\boldsymbol{v}_2^T\mathbb{X}_2^T(I-H)\boldsymbol{\varepsilon}\bigr]\bigr\}\\
&\quad\le C_6\bigl\{E\bigl[\operatorname{tr}\bigl(\mathbb{X}_2^T(I-H)\mathbb{X}_2\bigr)\operatorname{tr}\bigl(\boldsymbol{f}(\mathbf{U})^T(I-H)\boldsymbol{f}(\mathbf{U})\bigr)\bigr]\\
&\qquad+\sigma^2E\bigl[\operatorname{tr}\bigl(\mathbb{X}_2^T(I-H)\mathbb{X}_2\bigr)\bigr]\bigr\}\\
&\quad=O\bigl(n^2p_nq_n^{-2r}+np_n\bigr).
\end{aligned}$$

Thus we have

$$\mathrm{II}_{n3}=\nu_n\bigl(np_n^{1/2}q_n^{-r}+n^{1/2}p_n^{1/2}\bigr)O_P(1).$$

For II_{n4}, by $0<\gamma<1$,

$$\biggl(\sum_{j\notin\mathcal{A}}\|\nu_n\boldsymbol{v}_{2j}\|^\gamma\biggr)^{1/\gamma}\ge\biggl(\sum_{j\notin\mathcal{A}}\|\nu_n\boldsymbol{v}_{2j}\|^2\biggr)^{1/2}=\nu_n\|\boldsymbol{v}_2\|.$$

Accordingly,

$$\mathrm{II}_{n4}\ge\varphi_{n2}\nu_n^\gamma\|\boldsymbol{v}_2\|^\gamma.$$

By condition (A5)(b), for some $\|\boldsymbol{v}_2\|>0$, we have

$$\Pr\bigl(V_n(\boldsymbol{v}_1,\boldsymbol{v}_2)-V_n(\boldsymbol{v}_1,\mathbf{0})>0\bigr)\to1.$$

(ii) Let $\boldsymbol{\omega}_n$ be some $\sum_{j=1}^{k_n}d_j$-vector with $\|\boldsymbol{\omega}_n\|^2=1$. By Theorem 3.2(i), with probability tending to 1, we have the following result:

$$\frac{\partial Q_n(\boldsymbol{\beta}_{(1)})}{\partial\boldsymbol{\beta}_{(1)}}\bigg|_{\boldsymbol{\beta}_{(1)}=\hat{\boldsymbol{\beta}}_{(1)}}=\mathbb{X}_1^T(I-H)\mathbb{X}_1(\hat{\boldsymbol{\beta}}_{(1)}-\boldsymbol{\beta}_{10})-\mathbb{X}_1^T(I-H)\bigl(\boldsymbol{f}(\mathbf{U})+\boldsymbol{\varepsilon}\bigr)+\boldsymbol{\xi}_n=\mathbf{0},$$

where $\boldsymbol{\xi}_n=(\lambda_1\gamma\|\hat{\boldsymbol{\beta}}_1\|^{\gamma-2}\hat{\boldsymbol{\beta}}_1^T,\dots,\lambda_{k_n}\gamma\|\hat{\boldsymbol{\beta}}_{k_n}\|^{\gamma-2}\hat{\boldsymbol{\beta}}_{k_n}^T)^T$. We consider the limit distribution

$$\begin{aligned}
&n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\bigl[\mathbb{X}_1^T(I-H)\mathbb{X}_1\bigr](\hat{\boldsymbol{\beta}}_{(1)}-\boldsymbol{\beta}_{10})\\
&\quad=n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\mathbb{X}_1^T(I-H)\boldsymbol{f}(\mathbf{U})+n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\mathbb{X}_1^T(I-H)\boldsymbol{\varepsilon}\\
&\qquad-n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\boldsymbol{\xi}_n\\
&\quad\triangleq J_{n1}+J_{n2}+J_{n3}.
\end{aligned}$$

For J_{n1},

$$J_{n1}^2=n^{-1}\bigl|\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\mathbb{X}_1^T(I-H)\boldsymbol{f}(\mathbf{U})\bigr|^2=O_P\bigl(nq_n^{-2r}\bigr).$$

For J_{n3}, by conditions (A2) and (A4), we have

$$E\big(J_{n3}^2\big) \le n^{-1}\tau_1^{-1}\varphi_{n1}\gamma^2\sum_{j=1}^{k_n} E\|\hat{\boldsymbol{\beta}}_j\|^{2(\gamma-1)} = O\big(n^{-1}\varphi_{n1}k_n\big).$$

For J_{n2},

$$\begin{aligned} J_{n2} &= n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\mathbf{G}_1^T(I-H)\boldsymbol{\varepsilon} + n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\widetilde{\mathbb{X}}_1^T\boldsymbol{\varepsilon} \\ &\quad - n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\widetilde{\mathbb{X}}_1^T H\boldsymbol{\varepsilon} \\ &\triangleq K_{n1} + K_{n2} + K_{n3}. \end{aligned}$$

Under conditions (C1) and (C2),

$$\begin{aligned} EK_{n1}^2 &= n^{-1}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}E\big[\mathbf{G}_1^T(I-H)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T(I-H)\mathbf{G}_1\big]\Omega_{11}^{-1/2}\boldsymbol{\omega}_n \\ &= O\big(k_n q_n^{-2r}\big). \end{aligned}$$

By condition (A6), we have

$$\begin{aligned} EK_{n3}^2 &= n^{-1}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}E\big[\widetilde{\mathbb{X}}_1^T H\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T H\widetilde{\mathbb{X}}_1\big]\Omega_{11}^{-1/2}\boldsymbol{\omega}_n \\ &= O\big(n^{-1}k_n q_n\big). \end{aligned}$$

Now we focus on K_{n2}

$$K_{n2} = n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\widetilde{\mathbb{X}}_1^T\boldsymbol{\varepsilon} \triangleq \frac{1}{\sqrt{n}}\sum_{i=1}^n s_{ni}\varepsilon_i.$$

First,

$$\begin{aligned} &E(s_{ni}\varepsilon_i) = 0; \\ &\operatorname{Var}\left(\sum_{i=1}^n s_{ni}\varepsilon_i\right) = \sum_{i=1}^n \operatorname{Var}(s_{ni}\varepsilon_i) = \sigma^2. \end{aligned}$$

Next we verify the conditions of the Lindeberg-Feller central limit. For any $\epsilon > 0$,

$$\begin{aligned} \sum_{i=1}^n E\big[\big(s_{ni}^2\varepsilon_i^2\big)\mathbf{1}\big(|s_{ni}\varepsilon_i| > \epsilon\big)\big] &= nE\big[\big(s_{n1}^2\varepsilon_1^2\big)\mathbf{1}\big(|s_{n1}\varepsilon_1| > \epsilon\big)\big] \\ &\le n\big[E\big(s_{n1}^4\varepsilon_1^4\big)\big]^{1/2}\big[\Pr\big(|s_{n1}\varepsilon_1| > \epsilon\big)\big]^{1/2}. \end{aligned}$$

By condition (A6),

$$\begin{aligned} E\big(s_{n1}^4\varepsilon_1^4\big) &= n^{-2}E\big\{\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\big[\mathbf{x}_1 - E(\mathbf{x}_1|U_1)\big]\big[\mathbf{x}_1 - E(\mathbf{x}_1|U_1)\big]^T\Omega_{11}^{-1/2}\boldsymbol{\omega}_n\big\}^2 E\varepsilon_1^4 \\ &\le n^{-2}\rho_{\min}^2\big(\boldsymbol{\omega}_n\boldsymbol{\omega}_n^T\big)\rho_{\max}^2\big(\Omega_{11}^{-1}\big)E\big\{\big[\mathbf{x}_1 - E(\mathbf{x}_1|U_1)\big]^T\big[\mathbf{x}_1 - E(\mathbf{x}_1|U_1)\big]\big\}^2 E\varepsilon_1^4 \end{aligned}$$

$$\begin{aligned}
&\leq n^{-2}\rho_{\min}^{2}\bigl(\boldsymbol{\omega}_n\boldsymbol{\omega}_n^T\bigr)\rho_{\max}^{2}\bigl(\Omega_{11}^{-1}\bigr)E\varepsilon_1^4 k_n d^* \sum_{j=1}^{k_n}\sum_{k=1}^{d_j} E\bigl[X_{1jk}-E(X_{1jk}|U_1)\bigr]^4 \\
&= O\bigl(k_n^2 n^{-2}\bigr)
\end{aligned}$$

and

$$\begin{aligned}
P\bigl(|s_{n1}\varepsilon_1|>\epsilon\bigr) &\leq \frac{1}{\epsilon^2}E(s_{n1}\varepsilon_1)^2 \\
&= \frac{\sigma^2}{\epsilon^2}n^{-1}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}E\bigl\{\bigl[\mathbf{x}_1-E(\mathbf{x}_1|U_1)\bigr]\bigl[\mathbf{x}_1-E(\mathbf{x}_1|U_1)\bigr]^T\bigr\}\Omega_{11}^{-1/2}\boldsymbol{\omega}_n \\
&= \frac{\sigma^2}{\epsilon^2}n^{-1} = O\bigl(n^{-1}\bigr).
\end{aligned}$$

Thus we have

$$\sum_{i=1}^{n}E\bigl[\bigl(s_{ni}^2\varepsilon_i^2\bigr)\mathbf{1}\bigl(|s_{ni}\varepsilon_i|>\epsilon\bigr)\bigr] = O\bigl(nk_n n^{-1}n^{-1/2}\bigr) = o(1).$$

This means that $K_{n2}\xrightarrow{D}N(0,\sigma^2)$. Using Slutsky's theorem, we have

$$n^{-1/2}\boldsymbol{\omega}_n^T\Omega_{11}^{-1/2}\bigl[\mathbb{X}_1^T(I-H)\mathbb{X}_1\bigr](\hat{\boldsymbol{\beta}}_{(1)}-\boldsymbol{\beta}_{10})\xrightarrow{D}N\bigl(0,\sigma^2\bigr).$$

Let $u_n^2 = n^2\boldsymbol{\omega}_n^T(\mathbb{X}_1^T(I-H)\mathbb{X}_1)^{-1}\Omega_{11}(\mathbb{X}_1^T(I-H)\mathbb{X}_1)^{-1}\boldsymbol{\omega}_n$, then

$$n^{1/2}u_n^{-1}\boldsymbol{\omega}_n^T(\hat{\boldsymbol{\beta}}_{(1)}-\boldsymbol{\beta}_{10})\xrightarrow{D}N\bigl(0,\sigma^2\bigr). \qquad \square$$

Acknowledgements
This research was supported by the National Natural Science Foundation of China (Grant No. 11401340).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 5 January 2017 Accepted: 16 June 2017 Published online: 30 June 2017

References

1. Tibshirani, R: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267-288 (1996)
2. Frank, I, Friedman, J: A statistical view of some chemometrics regression tools. Technometrics **35**, 109-148 (1993)
3. Fan, J, Li, R: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348-1360 (2001)
4. Zou, H: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418-1429 (2006)
5. Zou, H, Hastie, T: Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**, 301-320 (2005)
6. Fu, W: Penalized regressions: the bridge versus the lasso. J. Comput. Graph. Stat. **7**, 397-416 (1998)
7. Knight, K, Fu, W: Asymptotics for lasso-type estimators. J. Comput. Graph. Stat. **28**, 1356-1378 (2000)
8. Liu, Y, Zhang, H, Park, C, Ahn, J: Support vector machines with adaptive l_q penalty. Comput. Stat. Data Anal. **51**, 6380-6394 (2007)
9. Huang, J, Horowitz, J, Ma, S: Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Stat. **36**, 587-613 (2008)
10. Wang, M, Song, L, Wang, X: Bridge estimation for generalized linear models with a diverging number of parameters. Stat. Probab. Lett. **80**, 1584-1596 (2010)

11. Chen, Z, Zhu, Y, Zhu, C: Adaptive bridge estimation for high-dimensional regression models. J. Inequal. Appl. **2016**, 258 (2016)
12. Huang, J, Ma, S, Xie, H, Zhang, C: A group bridge approach for variable selection. Biometrika **96**, 339-355 (2009)
13. Park, C, Yoon, Y: Bridge regression: adaptivity and group selection. J. Stat. Plan. Inference **141**, 3506-3519 (2011)
14. Engle, R, Granger, C, Rice, J, Weiss, A: Semiparametric estimates of the relation between weather and electricity sales. J. Am. Stat. Assoc. **81**, 310-320 (1986)
15. Xie, H, Huang, J: Scad-penalized regression in high-dimensional partially linear models. Ann. Stat. **37**, 673-696 (2009)
16. Donoho, D, Johnstone, I: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425-455 (1994)
17. Wang, L, Li, H, Huang, J: Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. J. Am. Stat. Assoc. **103**, 1556-1569 (2008)
18. Berndt, ER: The Practice of Econometrics: Classical and Contemporary. Addison-Wesley, Reading (1991)