Supplement for "The de-biased group Lasso estimation for varying coefficient models"

Toshio Honda

S.1 Technical proofs

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

We often appeal to the standard arguments based on Bernstein's inequality and reproduce the inequality from [S5] for reference.

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

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

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

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

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

we have

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

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

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

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

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

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

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

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

Proof of Lemma 1) Write

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

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

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

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

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

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

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

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

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

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

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

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

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

uniformly in l and j with probability tending to 1.

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

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

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

Proof of Lemma 2) Set

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

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

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

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

$$\mathrm{E}\{|B_{l_1}(Z_1)B_{l_2}(Z_1)X_{1,j_1}X_{1,j_2}|\} \le C_1 \quad \text{and}$$
$$\mathrm{E}\{|B_{l_1}(Z_1)B_{l_2}(Z_1)X_{1,j_1}X_{1,j_2}|^m\} \le \mathrm{E}\{|B_{l_1}(Z_1)X_{1,j_1}|^{2m}\} + \mathrm{E}\{|B_{l_2}(Z_1)X_{1,j_2}|^{2m}\} \le C_2L(C_3L)^{m-2}m!$$

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

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

with probability tending to 1 for some positive constant C_4.

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

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

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

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

This implies

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

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

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

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

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

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

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

The former set is related to the deviation condition and the latter one is related to the RE condition. According to Lemma 1 and the condition on λ_0, the probability of this intersection tends to 1.

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

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

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

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

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

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

This (S.12) is equivalent to

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

The above inequality yields

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

Note that (S.13) implies that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \in \Psi(\mathcal{S}_0, 3)$ since $\mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0}) = \mathrm{P}_1(\widehat{\boldsymbol{\beta}}_{\overline{\mathcal{S}}_0} - \boldsymbol{\beta}_{0\overline{\mathcal{S}}_0})$. Thus we recall

the definition of $\phi^2_{\widehat{\Sigma}}(\mathcal{S}_0, 3)$ and obtain

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

Finally by the RE condition, we have

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

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

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

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

Thus we have

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

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

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

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

Proof of Lemma 3) First we should evaluate

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

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

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

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

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

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

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

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

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

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

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

Proof of Lemma 5) Write

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

where $\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4$ are clearly defined in the last line. We evaluate $\widehat{D}_1, \widehat{D}_2, \widehat{D}_3, \widehat{D}_4$ uniformly in j and k. We suppress the subscripts j and k here.

$\widehat{D}_1$: Exactly as in the proof of Lemma 3, we have

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

uniformly in j and k with probability tending to 1 for some positive constant C_1.

$\widehat{D}_2$ and $\widehat{D}_3$: Recall the result in Proposition 2. Then the absolute value of the (a, b) element of $\widehat{D}_2$ is bounded from above by

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

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

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

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

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

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

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

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

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

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

Proof of Lemma 6) Write

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

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

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

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

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

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

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

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

Next we demonstrate (b). Since

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

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

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

The second result follows from (a), the first one, and (25).
$\|\widehat{A}_j\|_F$: The (a, b) element of $\widehat{A}_j$ is bounded from above by

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

Therefore

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

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

S.2 Additional numerical studies

S.2.1 Simulation studies

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

The models and the parameters such as n and p are the same as in Section 4. We used also the cv.gglasso function as well as in Section 4. We implemented the group SCAD by using the R package 'grpreg' version 3.2-1 (the cv.grpreg function). It is provided by Prof. Patrick Breheny. See [S1] for more details. Our weights of the adaptive group Lasso estimator are as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

S.2.2 A real data application

We applied the proposed de-biased group Lasso procedure to the Boston Housing data as in e.g. [S2] and [S4]. The data set is available in the R package 'MASS.' See also [S3] about the data set. The data set has 14 variables, crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, lstat, medv, and 506 samples. The details of these variables are given at the end of this section. We augmented the data set by adding some artificial variables.

In this study, we followed [S2] and [S4] and took Y = medv and lstat as the index variable. Note that [S2] does not deal with high-dimensional models. As for lstat, we defined Z as $Z = F(\text{lstat})$, where $F(\cdot)$ is the distribution function of $2\times$ the χ^2 distribution with d.f. 6. We did this transformation to make the distribution of Z close to that of the uniform distribution on $[0, 1]$. Note that [S2] and [S4] included only part of the original variables e.g. crim, rm, tax, and ptratio in their models. We removed only a dummy variable chas since it does seem to be significant in our preliminary analysis. The conditional number of the covariance matrix of 11 original variables exceeds 100. This setup is unfavorable to any data analysis procedure. The conditional number of the covariance matrix of only crim, rm, tax, and ptratio is about 14.

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

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

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

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

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

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

In the tables, $\|\widehat{b}_j\|^2$ and $\|\widetilde{\beta}_j\|^2$ are from the de-biased Lasso and the SCAD, respectively. We computed p-values in the tables in a similar way to the critical values in Section 4 by using Theorem 1. We tried $p = 100$ with $L = 5$ and the quadratic spline basis. The results of 24 larger $\|\widehat{b}_j\|^2$ are given in Table S.25. In [S4], they included only four original variables (rm, crim, tax, ptratio) and straightforward comparisons are very difficult.

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

necessary in practical situations.

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

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

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

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

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

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

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

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

References

[S1] P. Breheny. grpreg: Regularization paths for regression models with grouped covariates. 2019. R package version version 3.2-1.

[S2] Z. Cai and X. Xu. Nonparametric quantile estimators for dynamic smooth coefficient models. *J. Amer. Statist. Assoc.*, 103:1595–1608, 2008.

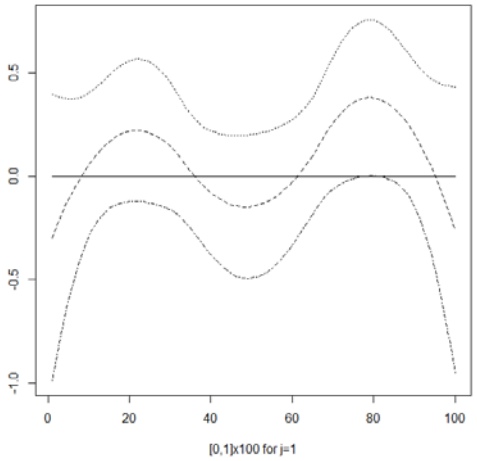
[S3] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Economics Managements*, 5:81-102, 1978.

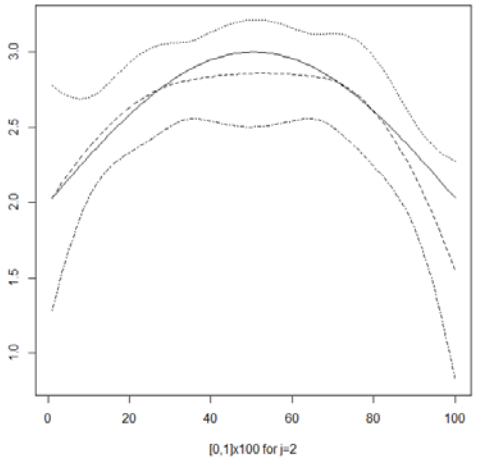
[S4] Y. Tang, X. Song, H. J. Wang, and Z. Zhu. Variable selection in high-dimensional quantile varying coefficient miodels. *J. Multivar. Anal.*, 122:115-132, 2013.

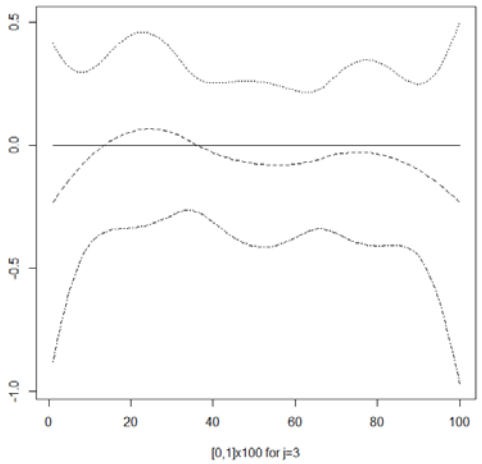
[S5] A. D. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

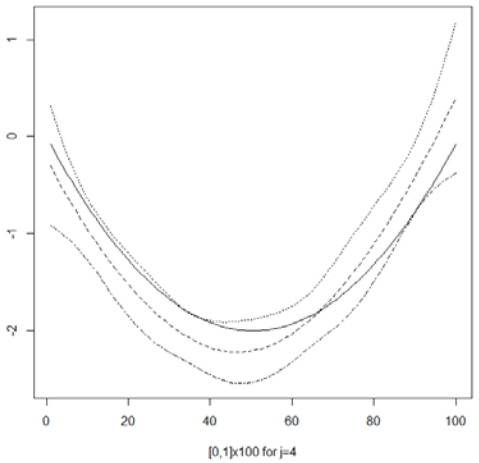
S.2.3 Confidence bands for g_j

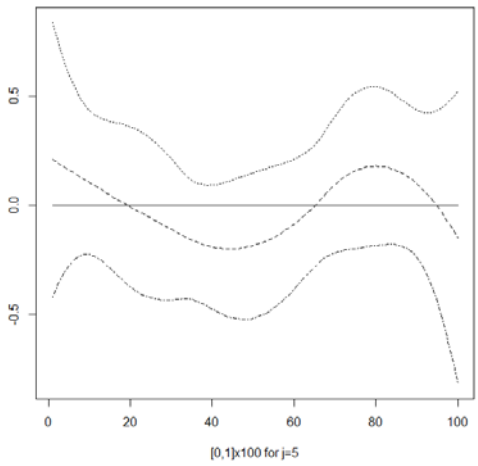
We present 8 figures of 95% confidence bands for g_j, j=1,⋯,8, and they are based on Theorem 1. We took one simulated sample for Model 1 with $p=n=350$. Real and broken lines represent ture g_j and estimated g_j, respectively. The other two lines are lower and upper bands for $g_j(t)$, not simultaneous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

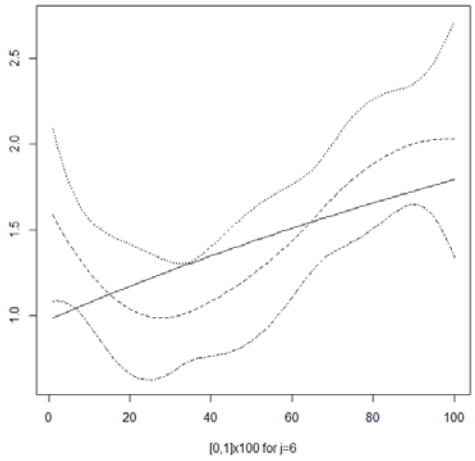


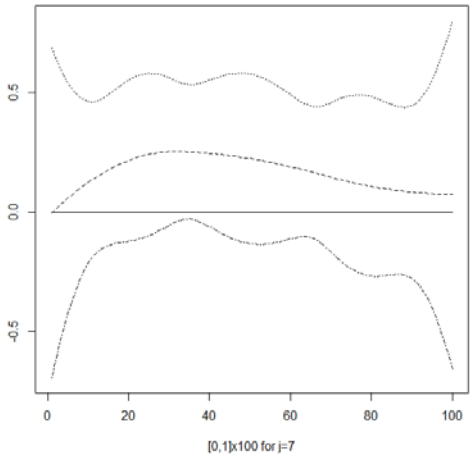


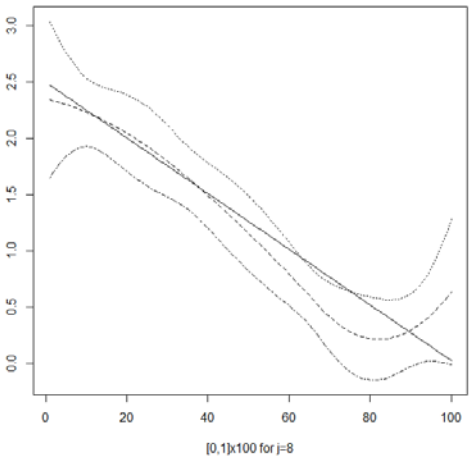





0.5
0.0
-0.5
0
20
40
60
80
100
[0,1]x100 for j=5


2.5
2.0
1.5
1.0
0
20
40
60
80
100
[0,1]x100 for j=6


0.5
0.0
-0.5
0
20
40
60
80
100
[0,1]x100 for j=7


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