Graduate School of Economics, Hitotsubashi University
Discussion Paper Series No. 2016-05

Variable selection and structure identification for varying coefficient Cox models

Toshio HONDA and Ryota YABE

First version：July 2016
Second version：September 2016
Third version：October 2016
This version：January 2017

Variable selection and structure identification for varying coefficient Cox models

Toshio Honda*

Graduate School of Economics, Hitotsubashi University, Kunitachi, Tokyo 186-8601, Japan

Ryota Yabe

Department of Economics, Shinshu University, Matsumoto, Nagano 390-8621, Japan

---

Abstract

We consider varying coefficient Cox models with high-dimensional covariates. We apply the group Lasso to these models and propose a variable selection procedure. Our procedure can cope with simultaneous variable selection and structure identification from a high dimensional varying coefficient model to a semivarying coefficient model. We also derive an oracle inequality and closely examine restrictive eigenvalue conditions. In this paper, we give the details for Cox models with time-varying coefficients. The theoretical results on variable selection can be easily extended to some other important models and we briefly mention those models since those models can be treated in the same way. The models considered in this paper are the most popular models among structured nonparametric regression models. The results of numerical studies are also reported.

Keywords: censored survival data, high-dimensional data, group Lasso, B-spline basis, structured nonparametric regression model, semivarying coefficient model
2010 MSC: 62G08, 62N01

---

1. Introduction

The Cox model is one of the most popular and useful models to analyze censored survival data. Since the Cox model was proposed in Cox[9], many authors

---

*Corresponding author
Email addresses: t.honda@r.hit-u.ac.jp (Toshio Honda),
ryotayabe@shinshu-u.ac.jp (Ryota Yabe)

have studied a lot of extensions or variants of the original Cox model to deal with complicated situations or carry out more flexible statistical analysis. In this paper, we consider varying coefficient models and additive models with high-dimensional covariates. These models with moderate numbers of covariates are investigated in many papers, for example, Huang et al.[18], Cai and Sun[8], and Cai et al.[7].

We apply the group Lasso (for example, see Lounici et al.[25] and Huang et al.[16]) to varying coefficient models with high-dimensional covariates to carry out variable selection and structure identification simultaneously. Although we focus on time-varying coefficient models here, our method can be applied to variable selection for another type of varying coefficient models and additive models and we briefly mention how to apply our procedure and how to derive the theoretical results.

Suppose that we observe censored survival times T_i and high-dimensional random covariates $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^T$. More specifically, we have n i.i.d. observations of

$$T_i = \min\{T_{0i}, C_i\}, \qquad \delta_i = I\{T_{0i} \leq C_i\}, \tag{1}$$

and p-dimensional covariate $\boldsymbol{X}_i(t)$ on the time interval $[0, \tau]$, where T_{0i} is an uncensored survival time and C_i is a censoring time satisfying the condition of the independent censoring mechanism as in section 6.2 of Kalbfleisch and Prentice[20]. Hereafter we set $\tau = 1$ for simplicity of presentation. Note that p can be very large compared to n in this paper, for example, $p = O(n^{c_p})$ for a very large positive constant c_p or $p = O(\exp(n^{c_p}))$ for a sufficiently small positive constant c_p. We assume that the standard setup for the Cox model holds as in chapter 5 of [20] and that T_i or $N_i(t) = I\{t \geq T_i\}$ has the following compensator $\Lambda_i(t)$ with respect to a suitable filtration $\{\mathcal{F}_t\}$:

$$d\Lambda_i(t) = Y_i(t) \exp\{\boldsymbol{X}_i^T(t)\boldsymbol{g}(t)\}\lambda_0(t)dt, \tag{2}$$

where $Y_i(t) = I\{t \leq T_i\}$, $\boldsymbol{g}(t) = (g_1(t), \ldots, g_p(t))^T$ is a vector of unknown functions on $[0, 1]$, $\boldsymbol{a}^T$ denotes the transpose of $\boldsymbol{a}$, and $\lambda_0(t)$ is a baseline hazard function. As in chapter 5 of [20], $\boldsymbol{X}_i(t)$ is predictable and

$$M_i(t) = N_i(t) - \Lambda_i(t) \tag{3}$$

is a martingale process with respect to $\{\mathcal{F}_t\}$. In the original Cox model, $\boldsymbol{g}(t)$ is a vector of unknown constants and we estimate this constant coefficient vector by maximizing the partial likelihood.

In this paper, we are interested in estimating $\boldsymbol{g}(t)$ in (2). Recently we have many cases where there are (ultra) high-dimensional covariates due to drastic development of data collecting technology. In such high-dimensional data, usually only a small part of covariates are relevant. However, we cannot directly apply standard or traditional estimating procedures to such high-dimensional data. Thus now a lot of methods for variable selection are available, for example, the SCAD and the Lasso. See Bühlmann and van de Geer[6] and Hastie et al.[14] for excellent reviews of these procedures for variable selection. See also Bickel et al.[3] and Zou[41] for the Lasso and the adaptive Lasso, respectively.

As for high dimensional Cox models with constant coefficient, Bradic et al.[4] studied the SCAD and Huang et al.[17], Kong and Nan[22], and Lemler[23] considered the Lasso. Zhang and Luo[36] proposed an adaptive Lasso estimator for the Cox model. The authors of [17] developed new ingenious techniques to derive oracle inequalities. We will fully use their techniques to derive our theoretical results such as an oracle inequality. Sun et al.[28] modified the Lasso penalty to incorporate side information. Wang et al.[32] proposed a hierarchical group penalty. Some variable screening procedures have also been proposed in Zhao and Li[40] and Yang et al.[34], to name just a few. Estimation of the baseline hazard function is considered in Guilloux et al.[13] in a high-dimensional setup. A model free screening procedure for censored data with high-dimensional covariates is proposed in Song et al.[27].

In this paper, we propose a group Lasso procedure to select relevant covariates and identify the covariates with constant coefficients among the relevant covariates, namely the true semivarying coefficient model from a much larger varying coefficient model. We can achieve this goal by a suitable two-stage procedure consisting of the proposed group Lasso with and an adaptively weighted Lasso procedure as in Yan and Huang[33] and Honda and Härdle[15] or the SCAD. In [33], the authors proposed an adaptive Lasso procedure for structure identification with no theoretical result. Our procedure can be applied to the varying coefficient model with an index variable $Z_i(t)$:

$$d\Lambda_i(t) = Y_i(t)\exp\{g_0(Z_i(t)) + \boldsymbol{X}_i^T(t)\boldsymbol{g}(Z_i(t))\}\lambda_0(t)dt \tag{4}$$

and the additive model:

$$d\Lambda_i(t) = Y_i(t)\exp\Big\{\sum_{j=1}^{p} g_j(X_{ij}(t))\Big\}\lambda_0(t)dt. \tag{5}$$

We mention these model later in section 4.

Some authors considered the same problem by using the SCAD. For example, see Lian et al.[24] and Zhang et al.[37]. They proved the existence of local optimizer satisfying the same convergence rate as ours. In contrast, we prove the existence of the global solution with desirable properties. In Bradic and Song[5], the authors applied penalties similar to ours to additive models and obtained theoretical results with model misspecifications considered. We have derived a better convergence rate in our cases. See Remark 3 in section 3 about the convergence rate. We also carefully examined the RE (restrictive eigenvalue) conditions. While the other authors considered the L_2 norm of the estimated second derivatives for additive models, we adopt the orthogonal decomposition approach to structure identification. We give some details on why we have adopted the orthogonal decomposition approach in Appendix C.

This paper is organized as follows. In section 2, we describe our group Lasso procedure for time-varying coefficient models. Then we present our theoretical results in section 3. We mention the two other models in section 4. The results of numerical studies are reported in section 5. The proofs of our theoretical results are postponed to section 6 and section 7 concludes this paper. We collected useful properties of our basis functions and the proofs of technical lemmas in Appendices A-D.

We define some notation and symbols here. In this paper, C, C_1, C_2, $\ldots$ are positive generic constants and their values change from line to line. For a vector $\boldsymbol{a}$, $|\boldsymbol{a}|$, $|\boldsymbol{a}|_1$, and $|\boldsymbol{a}|_\infty$ mean the L_2 norm, the L_1 norm, and the sup norm, respectively. For a function g on $[0, 1]$, $||g||$, $||g||_1$, and $||g||_\infty$ stand for the L_2 norm, the L_1 norm, and the sup norm, respectively. For a symmetric matrix A, we denote the minimum and maximum eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. Besides, sign(a) is the sign of a real number a and $a_n \sim b_n$ means there are positive constants C_1 and C_2 such that $C_1 < a_n/b_n < C_2$. We write $\overline{\mathcal{S}}$ for the complement of a set $\mathcal{S}$. For a function $g(z)$ and a constant c, $g(z) \equiv c$ and $g(z) \not\equiv c$ means that a function $g(z)$ is c and it is not a constant c, respectively.

2. Group Lasso procedure

First we decompose $g_j(t)$, $j = 1, \ldots, p$, into the constant part and the non-constant part:

$$g_j(t) = g_{cj} + g_{nj}(t), \tag{6}$$

where $\int_0^1 g_{nj}(t)dt = 0$. When $g_j(t) \not\equiv 0$, $g_j(t)$ is a non-zero constant or a non-constant function. We denote the index sets of relevant covariates by

$$\mathcal{S}_c = \{j \,|\, g_{cj} \neq 0\} \quad \text{and} \quad \mathcal{S}_n = \{j \,|\, g_{nj}(t) \not\equiv 0\} \tag{7}$$

and set

$$s_c = \#\mathcal{S}_c, \quad s_n = \#\mathcal{S}_n, \quad \text{and} \quad s_o = s_c + s_n,$$

where $\#A$ is the number of the elements of a set A. Even though p is large, only a small number of covariates are relevant in most cases. Then we consider sparse models and therefore we assume that $s_0 = o(L(\ln n)^{-1/2})$, where L is the dimension of our B-spline basis.

Next we introduce our spline basis $\overline{\boldsymbol{B}}(t)$ to approximate $g_j(t)$, $j = 1, \ldots, p$. We construct $\overline{\boldsymbol{B}}(t)$ from the L-dimensional equispaced B-spline basis $\boldsymbol{B}_0(t) = (b_{01}(t), \ldots, b_{0L}(t))^T$ on $[0, 1]$ and the basis has the following properties :

$$\overline{\boldsymbol{B}}(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_L(t) \end{pmatrix} = \begin{pmatrix} 1/\sqrt{L} \\ \boldsymbol{B}(t) \end{pmatrix} = A_0 \boldsymbol{B}_0(t) \quad \text{and} \quad \int_0^1 \overline{\boldsymbol{B}}(t)\overline{\boldsymbol{B}}^T(t)dt = L^{-1}I, \tag{8}$$

where

$$A_0 = \begin{pmatrix} \boldsymbol{a}_{01}^T \\ \boldsymbol{a}_{02}^T \\ \vdots \\ \boldsymbol{a}_{0L}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{1}^T/\sqrt{L} \\ A_{-1} \end{pmatrix}$$

and $\boldsymbol{1} = (1, \ldots, 1)^T$. Besides, we define $\boldsymbol{a}_{0j}$ and A_{-1} in the above equations. Note that for $j = 1, \ldots, L$,

$$b_j(t) = \boldsymbol{a}_{0j}^T \boldsymbol{B}_0(t)$$

and that $1/\sqrt{L}$ and $\boldsymbol{B}(T) = (b_2(t), \ldots, b_L(t))^T$ in (8) are designed for g_{cj} and $g_{nj}(t)$, respectively. Recall that $\boldsymbol{1}^T \boldsymbol{B}_0(t) \equiv 1$ and see Schumaker[26] for the definition of B-spline bases. We have collected how to construct $\overline{\boldsymbol{B}}(t)$ and A_0 and some useful properties of $\overline{\boldsymbol{B}}(t)$ and A_0 in Appendix A. We can use another basis which has desirable properties such as (A.1), (A.3), and (A.4) in Appendix A. We also use the local properties of the B-spline basis in the proofs.

We impose some technical assumptions on $\boldsymbol{g}(t)$.

Assumption G : $g_j(t),\ j = 1, \ldots, p$, are twice continuously differentiable and there is a positive constant C_g such that

$$\sum_{j=1}^{p} \|g_j\|_\infty \le C_g, \quad \sum_{j=1}^{p} \|g_j'\|_\infty \le C_g, \quad \text{and} \quad \sum_{j=1}^{p} \|g_j''\|_\infty \le C_g.$$

Besides we have

$$\min_{j \in \mathcal{S}_c} |g_{cj}| L^2 \to \infty \quad \text{and} \quad \min_{j \in \mathcal{S}_n} \|g_{nj}\| L^2 \to \infty.$$

Hereafter we take $L = c_L n^{1/5} (c_L > 0)$ for simplicity of presentation and the order of the B-spline basis should be larger than or equal to 2. In the former of Assumption G, most of g_j are irrelevant and satisfy $g_j(z) \equiv 0$ since we are dealing with sparse models. Then only a small number or bounded number of g_j such that only $j \in \mathcal{S}_c \cup \mathcal{S}_n$ are relevant in this assumption and the summations. The latter of Assumption G means relevant coefficient functions are larger than the spline approximation error. As for the identifiability of $\boldsymbol{g}(t)$, we need an assumption such as $\lambda_{\min}(\mathrm{E}\{\overline{\Sigma}\}) > C_1/L$ for a positive constant C_1, where $\mathrm{E}\{\overline{\Sigma}\}$ is defined in Proposition 3.

When Assumption G holds, there are $\gamma_j^* = (\gamma_{1j}^*, \gamma_{-1j}^{*T})^T \in R^L,\ j = 1, \ldots, p$, such that for a positive constant C_{approx} depending on C_g,

$$\sum_{j=1}^{p} \|g_j - \gamma_j^{*T} \overline{\boldsymbol{B}}(t)\|_\infty \le C_{approx} L^{-2}. \tag{9}$$

When $j \in \mathcal{S}_c$, we can take $\gamma_{1j}^* = \sqrt{L} g_{cj}$ and $\gamma_{-1j}^* \in R^{L-1}$ depends on $g_{nj}(t)$. If $j \in \overline{\mathcal{S}}_n$ and $j \in \overline{\mathcal{S}}_c$, we take $\gamma_{-1j}^* = 0$ and $\gamma_{1j}^* = 0$, respectively. See Appendix A for more details on these $\gamma_j^* = (\gamma_{1j}^*, \gamma_{-1j}^{*T})^T$.

We state assumptions on our Cox model before we describe the log partial likelihood for new covariates

$$\boldsymbol{W}_i(t) = \boldsymbol{X}_i(t) \otimes \overline{\boldsymbol{B}}(t), \tag{10}$$

where $\otimes$ means the Kronecker product.

Assumption M : $|X_{1j}(t)| \le C_X$ uniformly in j and t for a positive constant C_X. We also have $\mathrm{E}\{Y_1(1)\} \ge C_Y$ for a positive constant C_Y. Besides, the baseline hazard function is bounded from above and satisfies $\lambda_0(t) \ge C_\lambda$ on $[0, 1]$ for a positive constant C_λ.

The first one is used to evaluate the inside of the exponential function. When we deal with additive models, we can do without it. The other ones are standard in the literature.

We denote the log partial likelihood by $L_p(\gamma)$:

$$L_p(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \gamma^T \boldsymbol{W}_i(t)dN_i(t) - \int_0^1 \ln\Big[\sum_{i=1}^{n} Y_i(t)\exp\{\gamma^T \boldsymbol{W}_i(t)\}\Big]d\overline{N}(t), \quad (11)$$

where $\gamma = (\gamma_1^T, \dots, \gamma_p^T)^T \in R^{pL}$ and $\overline{N}(t) = n^{-1}\sum_{i=1}^{n} N_i(t)$. We also use the same sample mean notation for $M_i(t)$ and $Y_i(t)$.

Set

$$\ell_p(\gamma) = -L_p(\gamma) \quad (12)$$

for notational convenience. Then we should minimize this $\ell_p(\gamma)$ with respect to γ. However, when pL is larger than n, we cannot carry out this minimization properly and we add some penalty as in the literature on high-dimensional data. We define the convex penalty :

$$P_1(\gamma) = \sum_{j=1}^{p}(|\gamma_{1j}| + |\gamma_{-1j}|). \quad (13)$$

This $P_1(\gamma)$ also plays the role of the L_1 norm for $\gamma \in R^{pL}$ and is a very important technical tool in this paper. Besides, we define a kind of sup norm $P_\infty(\gamma)$ by

$$P_\infty(\gamma) = \max_{1\le j\le p} |\gamma_{1j}| \vee |\gamma_{-1j}|, \quad (14)$$

where $a \vee b = \max\{a, b\}$. This is also an important technical tool.

Thus our group Lasso objective function is defined by

$$Q_1(\gamma; \lambda) = \ell_p(\gamma) + \lambda P_1(\gamma). \quad (15)$$

Our group Lasso estimate is given by

$$\widehat{\gamma} = \operatorname*{argmin}_{\gamma\in R^{pL}} Q_1(\gamma; \lambda). \quad (16)$$

If we are interested in only variable selection, we should minimize

$$Q(\gamma; \lambda) = \ell_p(\gamma) + \lambda \sum_{j=1}^{p} |\gamma_j|. \quad (17)$$

By using the results on the Lasso for quantile regression in Belloni and Chernozhukov[1], Tang et al.[29] and Kato[21] considered variable selection for varying coefficient and additive quantile regression models, respectively.

We state our theoretical results only for $Q_1(\gamma;\lambda)$ in section 3 since we can deal with $Q(\gamma;\lambda)$ in the same way. However, the sup norm $P_\infty(\gamma)$ should be modified to $P_\infty(\gamma) = \max_{1\le j\le p} |\gamma_j|$ and the oracle inequality gives an upper bound of $\sum_{j=1}^{p} |\widehat{\gamma}_j - \gamma_j^*|$ when we deal with $Q(\gamma;\lambda)$.

The standard optimization theory implies that we have for $\widehat{\gamma}$ in (16),

$$\frac{\partial \ell_p}{\partial \gamma_j}(\widehat{\gamma}) = -\lambda \nabla_j P_1(\widehat{\gamma}), \quad j = 1, \dots, p, \tag{18}$$

where $\nabla_j P_1(\gamma)$ is the subgradient of $P_1(\gamma)$ with respect to γ_j and it consists of

$$\nabla_{1j}|\gamma_{1j}| = \begin{cases} \mathrm{sign}(\gamma_{1j}), & |\gamma_{1j}| \neq 0 \\ \epsilon_{1j}, & \gamma_{1j} = 0 \end{cases},$$

and

$$\nabla_{-1j}|\gamma_{-1j}| = \begin{cases} \gamma_{-1j}/|\gamma_{-1j}|, & |\gamma_{-1j}| \neq 0 \\ \boldsymbol{\epsilon}_{-1j}, & \gamma_{-1j} = 0 \end{cases}.$$

Note that $|\epsilon_{1j}| \le 1$ and $|\boldsymbol{\epsilon}_{-1j}| \le 1$ and ∇_{1j} and ∇_{-1j} are subgradients with respect to γ_{1j} and γ_{-1j}, respectively. See chapter 5 of [14] for more details about convex optimality conditions.

Consequently from (8), our estimates of g_{cj} and g_{nj} are

$$\widehat{g}_{cj} = \widehat{\gamma}_{1j}/\sqrt{L} \quad \text{and} \quad \widehat{g}_{nj}(t) = \boldsymbol{B}^T(t)\widehat{\gamma}_{-1j}. \tag{19}$$

Remark 1. The Lasso is not necessarily selection consistent although our simulation results are very good for variable selection. This phenomenon is closely examined in [41] and some other papers for L_2 linear regression. They proved that the Lasso needs a restrictive condition on covariates to be selection consistent even for L_2 linear regression. See section 2.7 in [6]. Hence the adaptive Lasso is proposed in [41]. Recently more general adaptively weighted Lasso procedures have been considered in the literature. Consequently the Lasso procedure often has to be followed by a next step like an adaptively weighted Lasso procedure or the SCAD. The SCAD also needs a good initial estimate or should have a smaller number of covariates. We usually calculate the weights of adaptively weighted Lasso procedures based on estimators with desirable properties such as so-called screening consistency. See section 2.8 in [6] about these kinds of two step procedures. For Cox models, the authors of [36] and [15] considered adaptively weighted Lasso procedures. See also Fan et al.[10] about L_1 regression and Fan et al.[11] for a more general principle. However, we have never seen our orthonormal basis applied to structure identification for Cox models. Our Theorem 1, which is an oracle inequality and a standard main result in the Lasso literature, gives a solid theoretical basis for our group Lasso procedure to be used as a first step for those adaptively weighted group Lasso or group SCAD procedures for Cox models. As one of the reviewers pointed out, we can use $Q(\gamma;\lambda)$ first and then apply $Q_1(\gamma;\lambda)$ with adaptive weights or the group SCAD. However, when we use $Q_1(\gamma;\lambda)$ first and $Q_1(\gamma;\lambda)$ with adaptive weights or the group SCAD next, we will be able to remove more of irrelevant non-constant components at the first step based on $Q_1(\gamma;\lambda)$. Inclusion of one irrelevant non-constant component increases the dimension of $\boldsymbol{W}_i(t)$ by $L-1$. Our theoretical results cover both strategies. From a theoretical point of view, if we choose a threshold value t_λ based on our theoretical results in section 3 and define $\widehat{\mathcal{S}_c}$ and $\widehat{\mathcal{S}_n}$ by

$$\widehat{\mathcal{S}_c} = \{j \,|\, |\widehat{g}_{cj}| > t_\lambda\} \quad \text{and} \quad \widehat{\mathcal{S}_n} = \{j \,|\, \|\widehat{g}_{nj}\| > t_\lambda\}, \tag{20}$$

they are consistent estimators of $\mathcal{S}_c$ and $\mathcal{S}_n$, respectively. Then we estimate the parameters based on $\widehat{\mathcal{S}_c}$ and $\widehat{\mathcal{S}_n}$. This is called the thresholded Lasso in the literature. See sections 2.9 and 7.6.2 in [6]. However, adaptively weighted Lasso procedures are much more popular in the literature. This is partly because the Lasso estimator is a biased one.

Remark 2. In some situations, we should assume

$$\mathcal{S}_n \subset \mathcal{S}_c. \tag{21}$$

We may incidentally have $g_{cj} = 0$ for $j \in \mathcal{S}_n$ even if $g_{nj}(z) \not\equiv 0$. This can happen partly because the value of g_{cj} can depend on the definition of the decomposition of $g_j(z)$ into g_{cj} and $g_{nj}(z)$. However, this will rarely happen and g_{cj} should be included into the model if the nonparametric regression coefficient function for j is included in the model. Then we can define a kind of hierarchical penalty $P_h(\gamma)$ as in (22) by taking the assumption in (21) into consideration and following Zhao et al.[39] and Zhao and Leng[38]. We take the subscript h of $P_h(\gamma)$ from the hierarchical assumption (21) and our hierarchical penalty $P_h(\gamma)$ is

$$P_h(\gamma) = \sum_{j=1}^{p} (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{1/q} + \sum_{j=1}^{p} |\gamma_{-1j}| \tag{22}$$

for some fixed $q > 1$. Then we can derive almost the same result for

$$\widehat{\gamma} = \operatorname*{argmin}_{\gamma \in R^{pL}} Q_h(\gamma; \lambda), \text{ where } Q_h(\gamma; \lambda) = \ell_p(\gamma) + \lambda P_h(\gamma),$$

as for $Q_1(\gamma; \lambda)$. When we deal with $Q_h(\gamma; \lambda)$, $P_1(\gamma)$ and $P_\infty(\gamma)$ still play the role of the L_1 and sup norms, respectively and the oracle inequality is an inequality about $P_1(\widehat{\gamma} - \gamma^*)$. We describe some more details in Appendix E in the supplement to this paper. When we assume (21) and the group Lasso based on $Q_1(\gamma; \lambda)$ concludes that $\|g_{nj}\| > 0$ and $|g_{cj}| = 0$, we may have to take (21) into consideration and modify this conclusion to the one that both of them are relevant for this j.

3. Oracle inequality

An oracle inequality for $\widehat{\gamma}$ from $Q_1(\gamma; \lambda)$ is given in Theorem 1. All the proofs are postponed to section 6. First we define some notation. We borrow some notation from [17] and proceed as in [17]. Some other notation is standard in the literature of the Cox model and the Lasso.

Let γ_S consist of $\{\gamma_{1j}\}_{j \in \mathcal{S}_c}$ and $\{\gamma_{-1j}\}_{j \in \mathcal{S}_n}$. On the other hand, $\gamma_{\overline{S}}$ consists of $\{\gamma_{1j}\}_{j \in \overline{\mathcal{S}}_c}$ and $\{\gamma_{-1j}\}_{j \in \overline{\mathcal{S}}_n}$.

We need some notation to give explicit expressions of the derivatives of $\ell_p(\gamma)$.

$$S^{(k)}(t, \gamma) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \boldsymbol{W}_i^{\otimes k}(t) \exp\{\boldsymbol{W}_i^T(t)\gamma\}, \tag{23}$$

where $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^T$. In addition,

$$\widetilde{\boldsymbol{W}}_n(t, \gamma) = \frac{S^{(1)}(t, \gamma)}{S^{(0)}(t, \gamma)} \quad \text{and} \quad V_n(t, \gamma) = \frac{S^{(2)}(t, \gamma)}{S^{(0)}(t, \gamma)} - (\widetilde{\boldsymbol{W}}_n(t, \gamma))^{\otimes 2}. \tag{24}$$

Hence we have the following expressions of the derivatives of $\ell_p(\gamma)$, which are denoted by $\dot{\ell}_p(\gamma)$ and $\ddot{\ell}_p(\gamma)$:

$$\frac{\partial \ell_p}{\partial \gamma}(\gamma) = -\frac{1}{n} \sum_{i=1}^{n} \int_0^1 \{\boldsymbol{W}_i(t) - \widetilde{\boldsymbol{W}}_n(t, \gamma)\} dN_i(t) = \dot{\ell}_p(\gamma) \tag{25}$$

and

$$\frac{\partial^2 \ell_p}{\partial \gamma \partial \gamma^T}(\gamma) = \int_0^1 V_n(t, \gamma) d\overline{N}(t) = \ddot{\ell}_p(\gamma) \tag{26}$$

Note that $\dot{\ell}_p(\gamma)$ and $\ddot{\ell}_p(\gamma)$ are define in the above equations.

In Proposition 1, we prove that $\widehat{\gamma}$ is in a restricted parameter space. We define some more notation to state Proposition 1. Set

$$D_\ell = P_\infty(\dot{\ell}_p(\gamma^*)) \quad \text{and} \quad \widehat{\boldsymbol{\theta}} = \widehat{\gamma} - \gamma^*. \tag{27}$$

We evaluate D_ℓ later in Proposition 2. We define $\boldsymbol{\theta}_{\mathcal{S}}$ and $\boldsymbol{\theta}_{\overline{\mathcal{S}}}$ in the same way as $\gamma_{\mathcal{S}}$ and $\gamma_{\overline{\mathcal{S}}}$. Recall that $\gamma^* = (\gamma_1^{*T}, \dots, \gamma_p^{*T})^T$ is given in (9). This proposition follows from only (18).

Proposition 1. *If $\lambda > D_\ell$, we have*

$$(\widehat{\gamma} - \gamma^*)^T\{\dot{\ell}_p(\widehat{\gamma}) - \dot{\ell}_p(\gamma^*)\} \leq (\lambda + D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}) - (\lambda - D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}})$$

and

$$(\lambda - D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}}) \leq (\lambda + D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}).$$

Therefore if $D_\ell \leq \xi\lambda$ ($\xi < 1$), we have

$$P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}}) \leq \frac{1+\xi}{1-\xi}P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}).$$

We define a restricted parameter space $\Theta(\zeta)$ by

$$\Theta(\zeta) = \{\boldsymbol{\theta} \in R^{pL} \mid P_1(\boldsymbol{\theta}_{\overline{\mathcal{S}}}) \leq \zeta P_1(\boldsymbol{\theta}_{\mathcal{S}})\}.$$

For $\boldsymbol{\theta} \in \Theta(\zeta)$, we have

$$P_1(\boldsymbol{\theta}) \leq (1+\zeta)P_1(\boldsymbol{\theta}_{\mathcal{S}}) \quad \text{and} \quad P_1(\boldsymbol{\theta}_{\mathcal{S}}) \leq s_0^{1/2}|\boldsymbol{\theta}_{\mathcal{S}}| \leq s_0^{1/2}|\boldsymbol{\theta}|. \tag{28}$$

Recall that s_0 is defined just after (7).

To state the compatibility and restrictive eigenvalue conditions, we define $\kappa(\zeta, \Sigma)$ and $RE(\zeta, \Sigma)$ for an n.n.d.(non-negative definite) matrix Σ with some modifications adapted to our setup.

$$\kappa(\zeta, \Sigma) = \inf_{\boldsymbol{\theta}\in\Theta(\zeta),\, \boldsymbol{\theta}\neq 0} \frac{s_0^{1/2}(\boldsymbol{\theta}^T\Sigma\boldsymbol{\theta})^{1/2}}{P_1(\boldsymbol{\theta}_{\mathcal{S}})} \quad \text{and} \quad RE(\zeta, \Sigma) = \inf_{\boldsymbol{\theta}\in\Theta(\zeta),\, \boldsymbol{\theta}\neq 0} \frac{(\boldsymbol{\theta}^T\Sigma\boldsymbol{\theta})^{1/2}}{|\boldsymbol{\theta}|}.$$

The latter is more commonly used in the literature of the Lasso. It is known that

$$\kappa^2(\zeta, \Sigma) \geq RE^2(\zeta, \Sigma) \geq \lambda_{\min}(\Sigma)$$

and that if $\Sigma_1 - \Sigma_2$ is n.n.d., we also have

$$\kappa(\zeta, \Sigma_1) \geq \kappa(\zeta, \Sigma_2) \quad \text{and} \quad RE(\zeta, \Sigma_1) \geq RE(\zeta, \Sigma_2).$$

Some more notation is necessary for Theorem 1. Set

$$C_W = 2C_X\{\lambda_{\max}(A_0A_0^T)\}^{1/2}, \quad RE^* = RE\Big(\frac{1+\xi}{1-\xi}, \ddot{\ell}_p(\gamma^*)\Big), \tag{29}$$

$$\kappa^* = \kappa\Big(\frac{1+\xi}{1-\xi}, \ddot{\ell}_p(\gamma^*)\Big), \quad \text{and} \quad \tau^* = \frac{s_0\lambda C_W}{(1-\xi)(\kappa^*)^2} \quad \text{for } \xi \in (0,1). \tag{30}$$

Note that C_W is bounded from above. We closely look at RE^* and κ^* in Proposition 3. Let η^* be the smaller solution of

$$\eta \exp(-\eta) = \tau^*$$

as in [17]. Note that τ^* should tend to 0 as in Remark 3. Actually it does for the choice of λ in Remark 3 due to our assumption on s_0.

We can deal with $Q(\gamma;\lambda)$ in (17) and $Q_h(\gamma;\lambda)$ in Remark 2 in almost the same way and drive the same results with just conformable changes.

Theorem 1. *Assume that Assumptions G and M hold. Then if $D_\ell \leq \xi\lambda$ for some $\xi \in (0,1)$, we have*

$$P_1(\widehat{\gamma} - \gamma^*) \leq \eta^*/C_W.$$

Then we also have

$$\max_{1\leq j\leq p} |\widehat{g}_{cj} - g_{cj}| \leq C_c\Big(\frac{\eta^*}{L^{1/2}} + L^{-2}\Big), \quad \max_{1\leq j\leq p} \|\widehat{g}_{nj} - g_{nj}\| \leq C_{n1}\Big(\frac{\eta^*}{L^{1/2}} + L^{-2}\Big),$$

$$\max_{1\leq j\leq p} \|\widehat{g}_{nj} - g_{nj}\|_\infty \leq C_{n2}\Big(\frac{\eta^*}{L^{1/2}} + L^{-2}\Big),$$

where C_c, C_{n1}, and C_{n2} depend on C_W, C_g, and the properties of the B-spline basis on $[0,1]$ *and they are bounded.*

Some remarks are in order.

Remark 3. When $p = O(n^{c_p})$ for some c_p, we have $D_\ell = O_p((n^{-1}\ln n)^{1/2})$ and should take $\lambda = C(n^{-1}\ln n)^{1/2}$ for some sufficiently large C. As in shown in Proposition 3, we usually have $(\kappa^*)^2 \sim L^{-1}$ with probability tending to 1 in suitable setups. Then when s_0 is bounded, $\tau^* \sim L(n^{-1}\ln n)^{1/2}$ and $\eta^*/\tau^* \to 1$. This leads to the convergence rate of $O(n^{-2/5}(\ln n)^{1/2})$ for $\widehat{g}_{cj}$ and $\widehat{g}_{nj}$. Our rate improves that of [5], which is $O(n^{-7/20}(\ln n)^{1/2})$ for their additive model in a similar setup. In their Theorems 1 and 2, $\lambda_n \geq C_1 n^{-1/4} d^{-1}(\ln n)^{1/2}$ for some positive constant C_1. Their convergence rate about coefficient estimaion has the order of $\{(n^{1/2}d)^{-1} \ln n\}^{1/2}$. Their d corresponds to our L, but it appears in the denominator, not in the numerator. If we take $d \sim n^{1/5}$ for this convergence rate, it reduces to $n^{-7/20}(\ln n)^{1/2}$. Our rate is optimal except for $(\ln n)^{1/2}$ for nonparametric regression under Assumption G when s_0 is bounded. Our results can deal with ultra high-dimensional cases if $p \sim \exp(n^{c_p})$ and c_p is sufficiently small. See Corollary 1 after Propositions 2 and 3.

Remark 4. Suppose that

$$\min_{j \in \mathcal{S}_c} |g_{cj}|/(n^{-2/5}(\ln n)^{1/2}) \to \infty \quad \text{and} \quad \min_{j \in \mathcal{S}_n} \|g_{nj}\|/(n^{-2/5}(\ln n)^{1/2}) \to \infty.$$

Then if we take t_λ satisfying $t_\lambda/\lambda \to \infty$ sufficiently slowly for λ in Remark 3, $\widehat{\mathcal{S}}_c$ and $\widehat{\mathcal{S}}_n$ in (20) are consistent estimators of $\mathcal{S}_c$ and $\mathcal{S}_n$, respectively. The conditions in this remark require that the relevant coefficients should be large enough. Note that $n^{-2/5}(\ln n)^{1/2}$ except for $(\ln n)^{1/2}$ comes from the optimal order of nonparametric regression under the assumption of the second order differentiability condition. When $p \sim \exp(n^{c_p})$, $\ln n$ should be replaced with $\ln p$.

Next we evaluate D_ℓ in Proposition 2, which is called the deviation condition. From Assumption M and application of Bernstein's inequality (for example, see van der Vaart and Wellner[31]), we have with probability larger than $1 - P_Y$,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(1) = \overline{Y}(1) > C_Y, \tag{31}$$

where

$$P_Y = \exp\Big\{-\frac{C_Y^2 n}{2(1 + 2C_Y/3)}\Big\}.$$

Since

$$\dot{\ell}_p(\gamma^*) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \{\boldsymbol{W}_i(t) - \widetilde{\boldsymbol{W}}_n(t, \gamma^*)\} dN_i(t), \tag{32}$$

we evaluate $\dot{\ell}_{op}$ defined in (33) and $\dot{\ell}_{op} - \dot{\ell}_p(\gamma^*)$ in (35). Note that this $\dot{\ell}_{op}$ has no argument.

$$\begin{aligned}\dot{\ell}_{op} &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big\{\boldsymbol{W}_i(t) - \frac{S_0^{(1)}(t)}{S_0^{(0)}(t)}\Big\} dN_i(t) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big\{\boldsymbol{W}_i(t) - \frac{S_0^{(1)}(t)}{S_0^{(0)}(t)}\Big\} dM_i(t),\end{aligned} \tag{33}$$

where

$$S_0^{(k)}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{W}_i^{\otimes k}(t)\exp\{\boldsymbol{g}^T(t)\boldsymbol{X}_i(t)\}, \quad k = 0, 1, 2. \tag{34}$$

$$\dot{\ell}_{op} - \dot{\ell}_p(\gamma^*) = \int_0^1 \Big\{\widetilde{\boldsymbol{W}}_n(t,\gamma^*) - \frac{S_0^{(1)}(t)}{S_0^{(0)}(t)}\Big\} d\overline{N}(t). \tag{35}$$

By combining evaluations of (33) and (35), we obtain Proposition 2. The proof is postponed to section 6. Recall that $\widetilde{\boldsymbol{W}}_n(t,\gamma^*)$ is defined in (24).

Proposition 2. *Assume that Assumptions G and M hold. Then we have*

$$P_\infty(\dot{\ell}_p(\gamma^*)) \le \frac{a_1}{L^{5/2}} + \frac{x(\ln n)^{1/2}}{\sqrt{n}}$$

with probability larger than

$$1 - P_Y - La_2\exp\{-a_3 nL^{-1}\} - 2pL\exp\Big\{-\frac{a_4x^2\ln n}{1 + x(n^{-1}L\ln n)^{1/2}}\Big\},$$

where a_j, $j = 1, \dots, 4$, *are positive constants depending only on the assumptions and they are independent of n.*

Finally we deal with κ^* and RE^*. In Proposition 3, we give their lower bounds. They are called the compatibility condition and the restricted eigenvalue condition, respectively.

Proposition 3. *Assume that Assumptions G and M hold. Then with probability larger than* $1 - P_Y - P_A - P_B - P_C$, *we have*

$$\begin{aligned}\kappa^2(\zeta, \ddot{\ell}_p(\gamma^*)) \ge{}& \exp(-C_X C_g)(1 + O(L^{-2}))\kappa^2(\zeta, \mathrm{E}\{\overline{\Sigma}\}) \\ &- s_0(1+\zeta)^2 L\Big\{\frac{c_1}{L^3} + \frac{x(\ln n)^{1/2}}{\sqrt{nL}}\Big\}\end{aligned}$$

and

$$RE^2(\zeta, \ddot{\ell}_p(\gamma^*)) \geq \exp(-C_X C_g)(1 + O(L^{-2}))RE^2(\zeta, \mathrm{E}\{\overline{\Sigma}\}) - s_0(1+\zeta)^2 L\Big\{\frac{c_2}{L^3} + \frac{x(\ln n)^{1/2}}{\sqrt{nL}}\Big\}$$

where

$$\overline{\Sigma} = \int_0^1 \overline{G}_Y(t)\lambda_0(t)dt, \quad \overline{G}_Y(t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)\{\boldsymbol{W}_i(t) - \boldsymbol{\mu}_Y(t)\}^{\otimes 2},$$

$$\boldsymbol{\mu}_Y(t) = \frac{\mathrm{E}\{Y_1(t)\boldsymbol{W}_1(t)\}}{\mathrm{E}\{Y_1(t)\}}, \quad P_A = 2(pL)^2 \exp\Big\{-\frac{c_3 x^2 \ln n}{1 + x(\ln n)^{1/2}(n^{-1}L)^{1/2}}\Big\},$$

$$P_B = 5(pL)^2 \exp\Big\{-\frac{c_4 x(n\ln n)^{1/2}}{1 + x^{1/2}(n^{-1}\ln n)^{1/4}}\Big\},$$

$$P_C = 2(pL)^2 \exp\Big\{-\frac{c_5 x^2 \ln n}{1 + x(n^{-1}\ln n)^{1/2}}\Big\}.$$

Note that c_j, $j = 1, \ldots, 5$, are positive constants depending only on the assumptions and they are independent of n.

In Propositions 2 and 3, the lower bounds of the probabilities depend on both p and n. By taking this p into consideration, we should choose x in the propositions to make the lower bounds of the probabilities tend to 1. When $p = O(n^{c_p})$ and $p \sim (\exp(n^{c_p}))$, we should take $x = C$ and $x = C\sqrt{\ln p/\ln n}$, respectively for a sufficiently large positive constant C in the propositions. See also the proof of Corollary 1 at the end of section 6 when $p \sim \exp(n^{c_p})$.

In the literature, it is often assumed that there is a positive constant C_1 such that $\lambda_{\min}(\mathrm{E}\{\overline{\Sigma}\}) \geq C_1/L$ due to (A.1) and (A.2) in Appendix A. Then for some positive constants C_2 and C_3, we have

$$\kappa^2(\zeta, \ddot{\ell}_p(\gamma^*)) \geq \frac{C_2}{L} + o_p(L^{-1}) \quad \text{and} \quad RE^2(\zeta, \ddot{\ell}_p(\gamma^*)) \geq \frac{C_3}{L} + o_p(L^{-1})$$

under the assumption of $s_0 = O(L(\ln n)^{-1/2})$ and $p = O(n^{c_p})$.

We give a corollary on unltra-high dimensional cases by employing Propositions 2 and 3. This corollary is proved at the end of section 6.

Corollary 1. *In addition to the assumptions in Theorem 1 and Propositions 2 and 3, we assume that $s_0 < C_a$ and $\lambda_{\min}(\mathrm{E}\{\overline{\Sigma}\}) \geq C_b/L$ for some positive constants C_a*

and C_b. Then if $p \sim \exp(n^{c_p})$ and $n^{c_p} = o(n^{2/5})$ for some positive constant c_p, we have

$$\eta^* \text{ in Theorem 1} \sim L(n^{-1} \ln p)^{1/2} \to 0$$

by taking $\lambda = C(n^{-1} \ln p)^{1/2}$ for some sufficiently large positive constant C.

4. Other models

4.1. Varying coefficient models with index variables

When we observe $(Z_i(t), \boldsymbol{X}_i(t))$ and $Z_i(t)$ is an influential variable treated as the index variable, the following model for the compensator is among candidates of our models for statistical analysis.

$$d\Lambda_i(t) = Y_i(t) \exp\{g_0(Z_i(t)) + \boldsymbol{X}_i^T(t)\boldsymbol{g}(Z_i(t))\}\lambda_0(t)dt, \tag{36}$$

where $Z_i(t) \in [0, 1]$, $\int_0^1 g_0(z)dz = 0$, and $g_j(z) = g_{cj} + g_{nj}(z)$, $j = 1, \dots, p$, as in section 2. Then we can proceed in almost the same way with

$$\boldsymbol{W}_i(t) = (\boldsymbol{B}^T(Z_i(t)), \boldsymbol{X}_i^T(t) \otimes \overline{\boldsymbol{B}}^T(Z_i(t)))^T,$$
$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{-10}^T, \gamma_{11}, \boldsymbol{\gamma}_{-11}^T, \dots, \gamma_{1p}, \boldsymbol{\gamma}_{-1p}^T)^T,$$
$$P_1(\boldsymbol{\gamma}) = \sum_{j=1}^{p} |\gamma_{1j}| + \sum_{j=0}^{p} |\boldsymbol{\gamma}_{-1j}|, \quad P_\infty(\boldsymbol{\gamma}) = \{\max_{1\le j\le p} |\gamma_{1j}| \vee |\boldsymbol{\gamma}_{-1j}|\} \vee |\boldsymbol{\gamma}_{-10}|,$$
$$Q_1(\boldsymbol{\gamma}; \lambda) = \ell_p(\boldsymbol{\gamma}) + \lambda P_1(\boldsymbol{\gamma}), \quad \text{and} \quad Q(\boldsymbol{\gamma}; \lambda) = \ell_p(\boldsymbol{\gamma}) + \lambda|\boldsymbol{\gamma}_{-10}| + \lambda \sum_{j=1}^{p} |\boldsymbol{\gamma}_j|.$$

We can define a hierarchical version $Q_h(\boldsymbol{\gamma})$ as in Remark 2.

We can carry out simultaneous variable selection and structure identification of this model as for time-varying coefficient models and we are able to prove the same results in almost the same way. Almost no change is necessary to the proofs of Proposition 1 and Theorem 1. When we consider Propositions 2 and 3, we should be a little careful in evaluating predictable variation processes and so on. Then we have to deal with terms like

$$n^{-1} \sum_{i=1}^{n} |b_{0j}(Z_i(t))|, \quad n^{-1} \sum_{i=1}^{n} |b_j(Z_i(t))|, \quad \text{and} \quad n^{-1} \sum_{i=1}^{n} |b_j(Z_i(t))b_k(Z_i(t))|$$

as compared to

$$|b_{0j}(t)|, \quad |b_j(t)|, \quad \text{and} \quad |b_j(t)b_k(t)|$$

for time-varying coefficient models. Note that we can use exponential inequalities for generalized U-statistics as given in Gine et al.[12] instead of Lemma 4.2 in [17] in the proof of Proposition 3. We give more details in Appendix D.

4.2. Additive models

When we have no specific index variable, the following additive model may be suitable.

$$d\Lambda_i(t) = Y_i(t)\exp\Big\{\sum_{j=1}^{p} g_j(X_{ij}(t))\Big\}\lambda_0(t)dt, \tag{37}$$

where $\int_0^1 g_j(x)dx = 0$ and $X_{ij}(t) \in [0,1]$. These $g_j(x)$ can be orthogonally decomposed into the linear part and the nonlinear part as well.

We should take $b_2(X_{ij}(t)) = (12L^{-1})^{1/2}(X_{ij}(t) - 1/2)$ and use $b_2(X_{ij}(t))$ and $(b_3(X_{ij}(t)), \ldots, b_L(X_{ij}(t)))^T$ for the linear part and the nonlinear part, respectively. We have no $b_1(X_{ij}(t))$ and divide $\boldsymbol{\gamma}_{-1j}$ into γ_{2j} and $\boldsymbol{\gamma}_{-2j} = (\gamma_{3j}, \ldots, \gamma_{Lj})^T$. Then we can apply the same group Lasso procedure for variable selection and structure identification with

$$\boldsymbol{W}_i(t) = (\boldsymbol{B}^T(X_{i1}(t)), \ldots, \boldsymbol{B}^T(X_{ip}(t)))^T, \quad \boldsymbol{\gamma}_{-1} = (\boldsymbol{\gamma}_{-11}^T, \ldots, \boldsymbol{\gamma}_{-1p}^T)^T,$$

$$P_1(\boldsymbol{\gamma}_{-1}) = \sum_{j=1}^{p}|\gamma_{2j}| + \sum_{j=1}^{p}|\boldsymbol{\gamma}_{-2j}|, \quad P_\infty(\boldsymbol{\gamma}_{-1}) = \max_{1\le j\le p}|\gamma_{2j}| \vee |\boldsymbol{\gamma}_{-2j}|,$$

$$Q_1(\boldsymbol{\gamma}_{-1};\lambda) = \ell_p(\boldsymbol{\gamma}_{-1}) + \lambda P_1(\boldsymbol{\gamma}_{-1}), \quad \text{and} \quad Q(\boldsymbol{\gamma}_{-1};\lambda) = \ell_p(\boldsymbol{\gamma}_{-1}) + \lambda\sum_{j=1}^{p}|\boldsymbol{\gamma}_{-1j}|.$$

We can define a hierarchical version $Q_h(\boldsymbol{\gamma}_{-1})$ as in Remark 2.

We have the same theoretical results with just conformable changes. We should be careful in the proofs of Propositions 2 and 3 as for varying coefficient models with index variables, too. We have to deal with terms like

$$n^{-1}\sum_{i=1}^{n}|b_{0j}(X_{i\ell}(t))|, \quad n^{-1}\sum_{i=1}^{n}|b_j(X_{i\ell}(t))|, \quad \text{and} \quad n^{-1}\sum_{i=1}^{n}|b_j(X_{i\ell}(t))b_k(X_{im}(t))|$$

as compared to

$$|b_{0j}(t)|, \quad |b_j(t)|, \quad \text{and} \quad |b_j(t)b_k(t)|$$

for time-varying coefficient models. We can use exponential inequalities for generalized U-statistics as given in Gine et al.[12] instead of Lemma 4.2 in [17] in the proof of Proposition 3.

5. Numerical studies

5.1. Simulation study

We carried out a simulation study for the two models in section 4 with the P_1 penalty because time-varying coefficient models are rather numerically intractable to us at present. We used the grpsurv function of the package 'grpreg' version 3.0-2 (Breheny[2]) for R in our numerical study and all the covariates are time-independent. We used R x64 3.3.1.

First we describe the data generating process of the covariates : $\{X_{ij}\}_{j=1}^{q}$, $\{X_{ij}\}_{j=q+1}^{p}$, and Z_i are mutually independent. Then X_{ij}, $j = q+1, \dots, p$, and Z_i follow $U(0,1)$ independently. We define $\{X_{ij}\}_{j=1}^{q}$ in (38).

$$X_{ij} = F(Y_{ij}), \quad j = 1, \dots, q, \tag{38}$$

where $\{Y_{ij}\}$ is a stationary Gaussian AR(1) process with $\rho = 0.3$ and $F(y)$ is the distribution function of Y_{ij}.

Next we gives the details for our varying coefficient model with an index variable Z. We took

$$\lambda_0(t) = 0.5, \quad g_1(z) = g_2(z) = 1, \quad g_3(z) = 4z, \quad g_4(z) = 4z^2.$$

The other functions are taken to be 0. Hence we have $s_c = 4$ and $s_n = 2$. Note that X_1 and X_2 are relevant for only the constant component and that X_3 and X_4 are relevant for both the constant component and the non-constant one. All the other covariates are irrelevant. We imposed no penalty on the coefficient vector for $g_0(z)$ in this simulation study. This does not affect the theoretical results. See the proof of Proposition 1. The censoring variable C_i follows the exponential distribution with mean= $1/0.85$ independently of all the other variables and the censoring rate is about 20%.

Then we describe the details for our additive model. We took

$$\begin{aligned} &\lambda_0(t) = 0.5, \quad g_1(x) = g_2(x) = 2^{1/2}(x - 1/2), \\ &g_3(x) = 2^{-1/2}\cos(2\pi x) + (x - 1/2), \quad g_4(x) = \sin(2\pi x). \end{aligned}$$

The other functions are taken to be 0. Hence we have $s_c = 4$ and $s_n = 2$ and note that X_1 and X_2 are relevant for only the linear component and that X_3 and X_4 are relevant for both the linear component and the nonlinear one. All the other covariates are irrelevant. The censoring variable C_i follows the exponential distribution with mean= $1/0.80$ independently of all the other variables and the censoring rate is about 30%.

When we carried out simulations, we took $n = 300$, $p = 500, 300, 150, 50$, $q = 8$. We took $L = 4$ and $L = 5$ for the varying coefficient model and the additive model, respectively. We used the quadratic spline basis and the repetition numbers are 400 for $p = 300, 150, 50$ and 100 for $p = 500$, respectively. The results are given in Tables 1 and 2. When $|\widehat{\gamma}_{1j}|$, $|\widehat{\gamma}_{-1j}|$, $|\widehat{\gamma}_{2j}|$, and $|\widehat{\gamma}_{-2j}|$ are less than 0.00001, they are put to 0. We give some figures of estimation errors of our procedure in Appendix F in the supplement. The Lasso estimator is a kind of biased estimator for variable selection. Thus our group Lasso estimator didn't perform very well in terms of estimation error.

In the tables, FNR, Correct, and FPR, respectively stand for

FNR: The rate of relevant covariates that are not chosen wrongly,
Correct: The rate of correct decisions,
FPR: The rate of irrelevant covariates that are wrongly chosen.

As for the tuning parameter λ, there is no theoretically definitive procedure and there is no result for selection consistency. In this simulation study, we chose λ by minimizing the AIC and the BIC. Our AIC and BIC for varying coefficient models are defined in (39) and (40).

$$AIC = \ell_p(\widehat{\gamma}) + \frac{1}{n}\{\widehat{s_c} + (L-1)\widehat{s_n}\}, \tag{39}$$

$$BIC = \ell_p(\widehat{\gamma}) + \frac{\ln n}{2n}\{\widehat{s_c} + (L-1)\widehat{s_n}\}, \tag{40}$$

where $\widehat{\gamma}$ is defined as in (16), $\widehat{s_c}$ is the number of non-zero $|\widehat{\gamma}_{1j}|$, and $\widehat{s_n}$ is the number of non-zero $|\widehat{\gamma}_{-1j}|$. Our AIC and BIC for additive models are similarly defined.

Tables 1 and 2 imply that the AIC minimization works very well. However, the BIC minimization does not work at all and we present only the tables of the AIC minimization here. Those of the BIC are given in Appendix F in the supplement. In Table 2, we sometimes missed the linear components of X_3 and X_4. If we incorporate the assumption in (21), we will not miss these linear components.

The results for the group SCAD are also given in Appendix F in the supplement. We took only $p = 50$ since the results for the other cases are unstable and bad. Probably the minimization of the grpsurv function does not work when we use it for the SCAD with large p and this may be a kind of general problem due to the nonconvexity of the SCAD penalty, not that of the specific R package. This is why various kinds of screening procedures have been proposed to give suitable initial values or reduce the numbers of covariates for the SCAD implementation.

Our procedure can be seen as a screening procedure as stated in Remark 1. To find the true model for large p, we go on to the second step for example, adaptively weighed group Lasso procedures or the SCAD procedure after our group Lasso procedure as the first step. Therefore it is very important to reduce the number of covariates properly. Our procedure based on $Q_1(\gamma; \lambda)$ can remove irrelevant non-constant or nonlinear components as shown, especially in Table 1. Irrelevant non-constant and nonlinear components will have serious negative effects on the dimension of Cox models at the second step since these components have larger dimensions than constant and linear components. Note again that we will not miss the linear components if we incorporate the assumption in (21) in Table 2.

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q (q = 8)$		X_{q+1} to X_p	
$p = 500$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.020	—	0.000	0.165	—	—	—	—
Correct	0.980	0.995	1.000	0.835	0.940	1.000	0.951	0.998
FPR	—	0.005	—	—	0.060	0.000	0.049	0.002
$p = 300$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.012	—	0.000	0.118	—	—	—	—
Correct	0.988	0.992	1.000	0.882	0.932	0.994	0.941	0.997
FPR	—	0.008	—	—	0.068	0.006	0.059	0.003
$p = 150$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.002	—	0.000	0.060	—	—	—	—
Correct	0.998	0.990	1.000	0.940	0.922	0.983	0.917	0.991
FPR	—	0.010	—	—	0.078	0.017	0.083	0.009
$p = 50$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.001	—	0.000	0.028	—	—	—	—
Correct	0.999	0.965	1.000	0.972	0.866	0.952	0.862	0.970
FPR	—	0.035	—	—	0.134	0.048	0.138	0.030

Table 1: Varying coefficient model with an index variable(AIC)

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q (q = 8)$		X_{q+1} to X_p	
$p = 500$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.010	—	0.405	0.075	—	—	—	—
Correct	0.990	0.990	0.595	0.925	1.000	0.998	0.999	0.993
FPR	—	0.010	—	—	0.000	0.002	0.001	0.007
$p = 300$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.014	—	0.335	0.038	—	—	—	—
Correct	0.986	0.986	0.665	0.962	0.999	0.990	0.999	0.988
FPR	—	0.014	—	—	0.001	0.010	0.001	0.012
$p = 150$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.011	—	0.232	0.018	—	—	—	—
Correct	0.989	0.966	0.768	0.982	0.996	0.978	0.997	0.975
FPR	—	0.034	—	—	0.004	0.022	0.003	0.025
$p = 50$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.001	—	0.122	0.004	—	—	—	—
Correct	0.999	0.896	0.878	0.996	0.986	0.907	0.987	0.916
FPR	—	0.104	—	—	0.014	0.093	0.013	0.084

Table 2: Additive model(AIC)

5.2. Real data analysis

We applied a varying coefficient model to the German Breast Cancer Study Group 2(GBSG2) dataset. The dataset is available from the package 'TH.data' for R. See https://cran.r-project.org/web/packages/TH.data/TH.data.pdf for more details on the data set.The data set consists of recurrence free survival times in days of 686 women with censoring indicators and eight covariates of three categorical covariates from tgrade to menostat and five continuous covariates from age to estrec. 56.5% of the observations were censored.

tgrade (X_1, X_2) : tumor grade, a ordered factor at levels I < II < III. $X_1 = tgrade2$ and $X_2 = tgrade3$ are dummy variables for II and III, respectively.

horTh (X_3) : hormonal therapy, a factor at two levels no and yes. X_3 is the dummy variable for yes.

menostat (X_4) : menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal). X_4 is the dummy variable for post

age (Z) : age of the patients in years

tsize (X_5) : tumor size (in mm)

pnodes (X_6) : number of positive nodes

progrec (X_7) : progesterone receptor (in fmol)

estrec (X_8) : estrogen receptor (in fmol).

We took age as Z as in [24]. Specifically,

$$Z = \Phi\Big(\frac{age - m_{age}}{v_{age}}\Big),$$

where m_{age} and v_{age} are the mean and the variance of age, respectively and $\Phi(x)$ is the distribution function of the standard normal distribution. Then our varying coefficient model has

$$g_0(Z_i) + \sum_{j=1}^{8} X_{ij} g_j(Z_i)$$

in the exponential function. Note that $g_0(z)$ has no constant component and we always included it in our selection with no penalty. We added $(p - 8)$ artificial covariates X_j with $g_j(z) \equiv 0$ for $j = 9, \dots, p$. Among the artificial covariates, X_9 and X_{10} take 0 and 1 and $X_{11}, \dots, X_{14}$ are continuous and correlated with tsize, pnodes, progrec,and estrec. The other ones are i.i.d. normal or uniform random variables. More details are given in Appendix G in the supplement. We considered two cases. We took $L = 4$ and $p = 500, 300, 150, 50$ and used the quadratic spline basis. We adopted the AIC minimization rule for tuning parameter selection as in our simulation study. Tables 4, 5, 7, and 8 report $|\widehat{g}_{cj}|$ and $\|\widehat{g}_{nj}\|$ by the SCAD and the coxph function of R with no additional artificial covariates. Entires in Z and menostat suggest that there seems to be serious multicollinearity among dummy variables.

Standardized case : We just standardized tsize, pnodes, progrec,and estrec by subtracting the mean and dividing the standard deviation.

The group Lasso selected for $p = 150$ and 300 the constant components of horTh, pnodes, progrec and no non-constant component, three components in total. When $p = 50$ and 500, it selected only the constant components of pnodes and progrec, two components in total. The group SCAD didn't work for $p = 150$, 300, or 500 even for $p = 50$ as shown in Table 3. Table 3 shows the numbers of false positive artificial covariates. Recall we always select $g_0(z)$ with no penalty on it.

Because of the consistency of the BIC for small and fixed p, the BIC results in Tables 4 and 7 suggest that screening procedures should select at least the following components.

Const : horTh, pnodes, progrec

Non-const: menostat

The constant component of tszie and the non-constant component of pnodes are relatively small for the AIC and disappeared for the BIC in Table 4.

The group Lasso missed the non-constant component of menostat for every p. We suspect from the results of Z and menostat in Tables 4 and 5 that that there is serious multicollinearity among dummy variables, tgrade2, tgrade3, horTh, and menostat. In addition, the design matrix also suggests the existence of it. See Appendix G in the supplement.

We think this multicollinearity is the reason that the group Lasso missed the menostat non-constant component. In order to recover variables such as this menostat, we should use another sure independence screening procedure simultaneously or closely look at the solution path. Aside from this menostat, our group Lasso procedure selected the necessary components for $p = 150$ and 300. For $p = 50$ and 500, the group Lasso missed horTh, too. Note that this horTh is not in the BIC result in Table 4, either.

Transformed case : We transformed tsize, pnodes, progrec,and estrec so that they are distributed on $[0, 1]$. The details are given in the supplement.

The group Lasso selected for every p the constant components of horTh, pnodes, progrec and no non-constant component, three in total. The group SCAD didn't work for $p = 150$, 300, or 500 and it selected constant components of tgrade2, tgrade3, horTh, pnodes, progrec and the non-constant component of menostat for $p = 50$, six in total. Table 6 shows the numbers of false positive artificial covariates.

Tables 7 and 8 suggest that screening procedures should select at least the constant components of horTh, pnodes, and progrec and the non-constant component of menostat as well. The group Lasso missed only the non-constant component of menostat for every p. The same comment for the standardized case applies to this menostat, too.

	X_9 and X_{10}		X_{11} to X_{14}		X_{15} to X_p	
	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
Lasso, p = 500	0	0	0	0	0	0
Lasso, p = 300	0	0	0	0	0	4
Lasso, p = 150	0	0	0	0	0	3
Lasso, p = 50	0	0	0	0	0	0
SCAD, p = 50	0	2	0	1	3	19

Table 3: False positive numbers (AIC, Standardized)

SCAD		Z	tgrate2	tgrade3	horTh	menostat	tsize	pnodes	progrec	estrec
AIC	Const.	—	0.602	0.724	0.428	0.000	0.031	0.304	0.409	0.000
	Non-Const.	1.063	0.228	0.000	0.000	1.761	0.000	0.040	0.000	0.000
BIC	Const.	—	0.000	0.000	0.000	0.000	0.000	0.304	0.478	0.000
	Non-Const.	1.127	0.000	0.000	0.000	1.755	0.000	0.000	0.000	0.000

Table 4: Norms by SCAD (No additional variables, Standardized)

coxph	Z	tgrate2	tgrade3	horTh	menostat	tsize	pnodes	progrec	estrec
Const.	—	0.613	0.724	0.437	3.311	0.120	0.264	0.412	0.158
Non-Const.	6.399	0.273	0.354	0.195	6.789	0.055	0.068	0.246	0.177

Table 5: Norms by coxph (No additional variables, Standardized)

	X_9 and X_{10}		X_{11} to X_{14}		X_{15} to X_p	
	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
Lasso, p = 500	0	0	0	0	9	0
Lasso, p = 300	0	0	0	0	5	0
Lasso, p = 150	0	0	0	0	1	0
Lasso, p = 50	0	0	0	0	0	0
SCAD, p = 50	0	1	0	3	0	6

Table 6: False positive numbers (AIC, Transformed)

SCAD		Z	tgrate2	tgrade3	horTh	menostat	tsize	pnodes	progrec	estrec
AIC	Const.	—	0.000	0.000	0.458	0.000	0.000	2.003	1.061	0.000
	Non-Const.	1.117	0.000	0.000	0.000	1.680	0.000	0.000	0.000	0.000
BIC	Const.	—	0.000	0.000	0.458	0.000	0.000	2.003	1.061	0.000
	Non-Const.	1.117	0.000	0.000	0.000	1.680	0.000	0.000	0.000	0.000

Table 7: Norms by SCAD (No additional variables, Transformed)

coxph	Z	tgrate2	tgrade3	horTh	menostat	tsize	pnodes	progrec	estrec
Const.	—	0.519	0.503	0.462	2.368	0.174	1.857	1.000	0.070
Non-Const.	5.007	0.333	0.432	0.202	5.225	0.073	0.246	0.389	0.312

Table 8: Norms by coxph (No additional variables, Transformed)

6. Proofs

We prove Propositions 1-3, Theorem 1, and Corollary 1. We present the proofs of technical lemmas in Appendix B.

For a vector $\boldsymbol{a}$ and a matrix A, $(\boldsymbol{a})_i$ and $(A)_{ij}$ mean the ith element of $\boldsymbol{a}$ and the (i, j) element of A, respectively.

Proof of Proposition 1. Note that

$$
\begin{aligned}
&(\widehat{\gamma} - \gamma^*)^T\{\dot{\ell}_p(\widehat{\gamma}) - \dot{\ell}_p(\gamma^*)\} \qquad (41)\\
= &\sum_{j\in\overline{\mathcal{S}}_n} \widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma}) + \sum_{j\in\mathcal{S}_n} \widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma})\\
&+ \sum_{j\in\overline{\mathcal{S}}_c} \widehat{\theta}_{1j}\frac{\partial \ell_p}{\partial \gamma_{1j}}(\widehat{\gamma}) + \sum_{j\in\mathcal{S}_c} \widehat{\theta}_{1j}\frac{\partial \ell_p}{\partial \gamma_{1j}}(\widehat{\gamma}) + \{-\widehat{\boldsymbol{\theta}}^T(\dot{\ell}_p(\gamma^*)\}\\
= &E_1 + E_2 + E_3 + E_4 + E_5 \geq 0
\end{aligned}
$$

Note that $E_k, k = 1, \dots, 5$, are defined in the above equation. The last inequality follows from the convexity of $\ell_p(\gamma)$ and we should recall that $\widehat{\boldsymbol{\theta}} = \widehat{\gamma} - \gamma^*$.

We evaluate E_k, $k = 1, \dots, 5$, by exploiting (18). Write $E_k = \sum_j E_{kj}$.

$\mathbf{E_1}$: Notice that $\widehat{\boldsymbol{\gamma}}_{-1j} = \widehat{\boldsymbol{\theta}}_{-1j}$. Then we should evaluate

$$E_{1j} = \widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma}) = \widehat{\boldsymbol{\gamma}}^T_{-1j}\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma}).$$

When $\widehat{\boldsymbol{\gamma}}_{-1j} \neq 0$, we have

$$E_{1j} = -\lambda|\widehat{\boldsymbol{\theta}}_{-1j}|. \qquad (42)$$

When $\widehat{\boldsymbol{\gamma}}_{-1j} = 0$, we have

$$E_{1j} = -\lambda|\widehat{\boldsymbol{\theta}}_{-1j}| = 0. \qquad (43)$$

From (42) and (43), we obtain

$$E_1 \leq -\lambda \sum_{j\in\overline{\mathcal{S}}_n} |\widehat{\boldsymbol{\theta}}_{-1j}|. \qquad (44)$$

$\mathbf{E_2}$: We should evaluate

$$E_{2j} = \widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma}).$$

We have $E_{2j} \le \lambda|\widehat{\boldsymbol{\theta}}_{-1j}|$ because

$$|\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma})| \le \lambda.$$

Thus we obtain

$$E_2 \le \lambda \sum_{j \in \mathcal{S}_n} |\widehat{\boldsymbol{\theta}}_{-1j}|. \tag{45}$$

E$_3$ and E$_4$: In a similar way, we obtain

$$E_3 \le -\lambda \sum_{j \in \overline{\mathcal{S}}_c} |\widehat{\theta}_{1j}| \quad \text{and} \quad E_4 \le \lambda \sum_{j \in \mathcal{S}_c} |\widehat{\theta}_{1j}|. \tag{46}$$

E$_5$: We have

$$E_5 \le P_1(\widehat{\boldsymbol{\theta}})D_\ell = (P_1(\widehat{\boldsymbol{\theta}_{\mathcal{S}}}) + P_1(\widehat{\boldsymbol{\theta}_{\overline{\mathcal{S}}}}))D_\ell. \tag{47}$$

(44), (45), (46), and (47) yield that

$$E_1 + E_2 + E_3 + E_4 + E_5 \le (\lambda + D_\ell)P_1(\widehat{\boldsymbol{\theta}_{\mathcal{S}}}) - (\lambda - D_\ell)P_1(\widehat{\boldsymbol{\theta}_{\overline{\mathcal{S}}}}).$$

The first and second inequalities follow from (41) and the above inequality. The third inequality follows from the following expression of the second one.

$$P_1(\widehat{\boldsymbol{\theta}_{\overline{\mathcal{S}}}}) \le \frac{\lambda + D_\ell}{\lambda - D_\ell} P_1(\widehat{\boldsymbol{\theta}_{\mathcal{S}}})$$

Hence the proof of the proposition is complete.

We establish the oracle inequality.

Proof of Theorem 1. First we define $D(\boldsymbol{\theta})$ by

$$D(\boldsymbol{\theta}) = \max_{i,j} \max_{0 \le t \le 1} |\boldsymbol{\theta}^T \boldsymbol{W}_i(t) - \boldsymbol{\theta}^T \boldsymbol{W}_j(t)|.$$

We need two lemmas.

Lemma 1.

$$D(\boldsymbol{\theta}) \le C_W P_1(\boldsymbol{\theta})$$

Lemma 2.

$$e^{-D(\boldsymbol{\theta})} \boldsymbol{\theta}^T \ddot{\ell}_p(\gamma^*) \boldsymbol{\theta} \le (\gamma^* + \boldsymbol{\theta} - \gamma^*)^T (\dot{\ell}_p(\gamma^* + \boldsymbol{\theta}) - \dot{\ell}_p(\gamma^*) \le e^{D(\boldsymbol{\theta})} \boldsymbol{\theta}^T \ddot{\ell}_p(\gamma^*) \boldsymbol{\theta}$$

Now we begin to prove the oracle inequality. If $\widehat{\boldsymbol{\theta}} = 0$, the desired inequality holds. Hence we assume $\widehat{\boldsymbol{\theta}} \neq 0$ and set

$$\widehat{\boldsymbol{b}} = \frac{\widehat{\boldsymbol{\theta}}}{P_1(\widehat{\boldsymbol{\theta}})}.$$

We have from Proposition 1 and the definition of $P_1(\boldsymbol{\gamma})$ that

$$\widehat{\boldsymbol{b}} \in \Theta\Big(\frac{1+\xi}{1-\xi}\Big) \quad \text{and} \quad P_1(\widehat{\boldsymbol{b}}) = P_1(\widehat{\boldsymbol{b}}_S) + P_1(\widehat{\boldsymbol{b}}_{\overline{S}}) = 1. \tag{48}$$

When $D_\ell \leq \xi\lambda$, the first inequality of Proposition 1 implies that the following inequalities hold at $x = 0$ and $x = P_1(\widehat{\boldsymbol{\theta}})$.

$$\widehat{\boldsymbol{b}}^T\{\dot{\ell}_p(\boldsymbol{\gamma}^* + x\widehat{\boldsymbol{b}}) - \dot{\ell}_p(\boldsymbol{\gamma}^*)\} \tag{49}$$
$$\leq (1+\xi)\lambda P_1(\widehat{\boldsymbol{b}}_S) - (1-\xi)\lambda P_1(\widehat{\boldsymbol{b}}_{\overline{S}})$$
$$= 2\lambda P_1(\widehat{\boldsymbol{b}}_S) - \lambda(1-\xi) \leq \frac{\lambda}{1-\xi}\{P_1(\widehat{\boldsymbol{b}}_S)\}^2. \tag{50}$$

We also used (48) here.

Note that (49) is monotone increasing and continuous in x due to the convexity of $\ell_p(\boldsymbol{\gamma})$ and we have (50) on $[0, P_1(\widehat{\boldsymbol{\theta}})]$. Let x_b be the maximum of x satisfying

$$\widehat{\boldsymbol{b}}^T\{\dot{\ell}_p(\boldsymbol{\gamma}^* + x\widehat{\boldsymbol{b}}) - \dot{\ell}_p(\boldsymbol{\gamma}^*)\} \leq \frac{\lambda}{1-\xi}\{P_1(\widehat{\boldsymbol{b}}_S)\}^2 \tag{51}$$

for any $s \in [0, x]$.

If we find an upper bound of x_b, say x_0, we have $P_1(\widehat{\boldsymbol{\theta}}) \leq x_0$. Therefore we will find an upper bound of x_b as in [17].

From Lemmas 1 and 2, we have for $\boldsymbol{\theta} = x\widehat{\boldsymbol{b}}$,

$$x\widehat{\boldsymbol{b}}^T\{\dot{\ell}_p(\boldsymbol{\gamma}^* + x\widehat{\boldsymbol{b}}) - \dot{\ell}_p(\boldsymbol{\gamma}^*)\} \geq x^2\exp\{-D(x\widehat{\boldsymbol{b}})\}\widehat{\boldsymbol{b}}^T\ddot{\ell}_p(\boldsymbol{\gamma}^*)\widehat{\boldsymbol{b}} \tag{52}$$
$$\geq x^2\exp\{-C_W x\}\widehat{\boldsymbol{b}}^T\ddot{\ell}_p(\boldsymbol{\gamma}^*)\widehat{\boldsymbol{b}}.$$

The definition of κ^* and (52) imply that

$$\widehat{\boldsymbol{b}}^T\{\dot{\ell}_p(\boldsymbol{\gamma}^* + x\widehat{\boldsymbol{b}}) - \dot{\ell}_p(\boldsymbol{\gamma}^*)\} \geq x\exp\{-C_W x\}\frac{(\kappa^*)^2}{s_0}\{P_1(\widehat{\boldsymbol{b}}_S)\}^2. \tag{53}$$

It follows from (51) and (53) that

$$\frac{\lambda s_0 C_W}{(1-\xi)(\kappa^*)^2} = \tau^* \geq C_W x\exp\{-C_W x\}.$$

Consequently we have from the definition of η^* and the above inequality that

$$C_W x_b \le \eta^* \quad \text{and} \quad \frac{\tau^*}{\eta^*} \to 1 \text{ if } \tau^* \to 0.$$

We have found that η^*/C_W is an upper bound of x_b and that $P_1(\widehat{\boldsymbol{\theta}}) \le \eta^*/C_W$.

As for the the rest of the theorem, the result on $\widehat{g}_{cj}$ is straightforward from (19). The upper bounds on $\widehat{g}_{nj}(t)$ follow from (A.1), (A.4), and the following inequalities.

$$|(\widehat{\gamma}_{-1j} - \gamma^*_{-1j})^T \boldsymbol{B}(t)| \le \{\lambda_{\max}(A_{-1}A_{-1}^T)\}^{1/2} |\widehat{\gamma}_{-1j} - \gamma^*_{-1j}||\boldsymbol{B}_0(t)| \quad \text{and}$$
$$|\boldsymbol{B}_0(t)| \le 1$$

Recall that the properties of our basis are collected in Appendix A.

Hence the proof of the theorem is complete.

Now we prove Proposition 2.

Proof of Proposition 2. We implicitly carry out our evaluation on $\{\overline{Y}(1) > C_Y\}$. $C_1, C_2, \ldots$ are generic positive constants and they depend only on the assumptions.

First we deal with (35), which is represented as

$$\int_0^1 \Big[\frac{S_0^{(0)}(t)\{S^{(1)}(t,\gamma^*) - S_0^{(1)}(t)\}}{S^{(0)}(t,\gamma^*)S_0^{(0)}(t)} + \frac{S_0^{(1)}(t)\{S_0^{(0)}(t) - S^{(0)}(t,\gamma^*)\}}{S^{(0)}(t,\gamma^*)S_0^{(0)}(t)}\Big] d\overline{N}(t). \tag{54}$$

We can rewrite the expression in (54) as

$$\begin{aligned}(54) &= (I \otimes A_0) \int_0^1 \Big[\frac{S_0^{(0)}(t)\{\overline{S}^{(1)}(t,\gamma^*) - \overline{S}_0^{(1)}(t)\}}{S^{(0)}(t,\gamma^*)S_0^{(0)}(t)} \\ &\quad + \frac{\overline{S}_0^{(1)}(t)\{S_0^{(0)}(t) - S^{(0)}(t,\gamma^*)\}}{S^{(0)}(t,\gamma^*)S_0^{(0)}(t)}\Big] d\overline{N}(t) \\ &= (I \otimes A_0)\Delta\dot{\ell}_p,\end{aligned} \tag{55}$$

where $\Delta\dot{\ell}_p$ is defined in the above equation,

$$\overline{S}^{(1)}(t,\gamma) = \frac{1}{n}\sum_{i=1}^n Y_i(t)(\boldsymbol{X}_i(t) \otimes \boldsymbol{B}_0(t)) \exp\{\boldsymbol{W}_i^T(t)\gamma\},$$
$$\overline{S}_0^{(1)}(t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)(\boldsymbol{X}_i(t) \otimes \boldsymbol{B}_0(t)) \exp\{\boldsymbol{X}_i(t)^T \boldsymbol{g}(t)\}.$$

Due to the definition of γ^*, we have uniformly in t and $k (0 \le k < p)$,

$$|S_0^{(0)}(t) - S^{(0)}(t,\gamma^*)| \le C_1 L^{-2},\ C_2 \le S_0^{(0)}(t) \wedge S^{(0)}(t,\gamma^*),\ S_0^{(0)}(t) \vee S^{(0)}(t,\gamma^*) \le C_3,$$

$$|(\overline{S}_0^{(1)}(t) - \overline{S}^{(1)}(t,\gamma^*))_{kL+j}| \le C_4 L^{-2} |b_{0j}(t)|,$$
$$|(\overline{S}_0^{(1)}(t))_{kL+j}| \vee |(\overline{S}^{(1)}(t,\gamma^*))_{kL+j}| \le C_5 |b_{0j}(t)|.$$

Now we evaluate $\Delta\dot{\ell}_p$. Its $(kL+j)$th element is bounded from above by

$$C_6 L^{-2} \int_0^1 |b_{0j}(t)| d\overline{N}(t). \tag{56}$$

for some positive constant C_6. First notice that

$$\int_0^1 |b_{0j}(t)| d\overline{N}(t) = \int_0^1 |b_{0j}(t)| d\overline{M}(t) + O(L^{-1}) \tag{57}$$

uniformly in j. This is bacause

$$d\overline{N}(t) - d\overline{M}(t) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \exp\{\boldsymbol{X}_i^T(t) \boldsymbol{g}(t)\} \lambda_0(t).$$

Then application of an exponential inequality for martingales (Lemma 2.1 in van de Geer[30]) yields

$$\Pr\Big(\max_{2 \le j \le L} \int_0^1 |b_{0j}(t)| d\overline{M}(t) > \frac{x}{L} \Big) \le L C_7 \exp\Big\{ -C_8 \frac{nL^{-1}x^2}{1+x} \Big\}. \tag{58}$$

We used the properties of the support of the B-spline basis in (57) and (58). Taking $x = 1$ in (58), we have established

$$|\Delta\dot{\ell}_p|_\infty \le \frac{C_9}{L^3} \tag{59}$$

with probability larger than $1 - LC_7 \exp\Big\{ -2^{-1} C_8 n L^{-1} \Big\}$. Recall $\Delta\dot{\ell}_p$ is defined in (55).

From (55), (59), and (A.3), we obtain

$$P_\infty(\dot{\ell}_{op} - \dot{\ell}_p(\gamma^*)) \le C_{10} L^{-5/2} \tag{60}$$

with probability larger than $1 - LC_7 \exp\Big\{ - 2^{-1} C_8 n L^{-1} \Big\}$. See (33) and (35) about $\dot{\ell}_{op}$.

Finally we deal with (33) by exploiting the same exponential inequality for martingales.

For the $(kL + j)$th element with $j = 1$, we have

$$\Pr\Big(|(\dot{\ell}_{op})_{kL+j}| \geq \frac{x(\ln n)^{1/2}}{\sqrt{nL}} \Big) \leq 2 \exp\Big\{ - \frac{C_{11} x^2 \ln n}{x(n^{-1} \ln n)^{1/2} + 1} \Big\}. \tag{61}$$

For the $(kL + j)$th element with $j \geq 2$, we have

$$\Pr\Big(|(\dot{\ell}_{op})_{kL+j}| \geq \frac{x(\ln n)^{1/2}}{\sqrt{nL}} \Big) \leq 2 \exp\Big\{ - \frac{C_{12} x^2 \ln n}{x(n^{-1} L \ln n)^{1/2} + 1} \Big\}. \tag{62}$$

We used the fact that

$$\int_0^1 b_j^2(t) \lambda_0(t) dt \leq C_\lambda \boldsymbol{a}_{0j}^T \Omega_0 \boldsymbol{a}_{0j} = O(L^{-1}) \tag{63}$$

when we evaluated the predictable variation process.

It follows from (61) and (62), that

$$P_\infty(\dot{\ell}_{op}) \leq x(\ln n)^{1/2} n^{-1/2} \tag{64}$$

with probability larger than

$$1 - 2pL \exp\Big\{ - \frac{C_{13} x^2 \ln n}{x(n^{-1} L \ln n)^{1/2} + 1} \Big\}. \tag{65}$$

Hence the desired result follows from (31), (60), and (64) and the proof of the proposition is complete.

We give the proof of Proposition 3.

Proof of Proposition 3. $C_1, C_2, \ldots$ are generic positive constants and they depend only on the assumptions. We use the following lemma, which is a version of Lemma 4.1(ii) in [17].

Lemma 3.

$$\kappa^2(\zeta, \Sigma_1) \geq \kappa^2(\zeta, \Sigma_2) - s_0(1 + \zeta)^2 L \max_{j,k} |(\Sigma_1 - \Sigma_2)_{jk}|$$

$$RE^2(\zeta, \Sigma_1) \geq RE^2(\zeta, \Sigma_2) - s_0(1 + \zeta)^2 L \max_{j,k} |(\Sigma_1 - \Sigma_2)_{jk}|$$

When $\Sigma_2 - \Sigma_1$ is n.n.d., we can replace $\Sigma_1 - \Sigma_2$ in the above inequalities with Δ such that $\Delta - (\Sigma_2 - \Sigma_1)$ is n.n.d.

We implicitly carry out our evaluation on $\{\overline{Y}(1) > C_Y\}$. First we outline the proof and then give the details. Recall that $V_n(t,\gamma)$, $S_0^{(0)}(t)$, and $S^{(0)}(t,\gamma)$ are defined in (24), (34), and (23), respectively.

Define $\widetilde{\Sigma}_0$ by

$$\widetilde{\Sigma}_0 = \int_0^1 V_n(t,\gamma^*)S_0^{(0)}(t)\lambda_0(t)dt \tag{66}$$

and set

$$\Delta_1 = \ddot{\ell}_p(\gamma^*) - \widetilde{\Sigma}_0 = \int_0^1 V_n(t,\gamma^*)d\overline{M}(t). \tag{67}$$

We treat Δ_1 by using the exponential inequalities for martingales.

Next define $\widetilde{\Sigma}$ by

$$\widetilde{\Sigma} = \int_0^1 V_n(t,\gamma^*)S^{(0)}(t,\gamma^*)\lambda_0(t)dt$$

and set $\Delta_2 = \widetilde{\Sigma}_0 - \widetilde{\Sigma}$. Since

$$|\boldsymbol{W}_i^T(t)\gamma^* - \boldsymbol{X}_i^T(t)\boldsymbol{g}(t)| \leq C_X C_{approx} L^{-2}$$

and we can use the results on predictable variation process in evaluating Δ_1, we can easily prove

$$\max_{j,k} |(\Delta_2)_{jk}| \leq C_1 L^{-3}. \tag{68}$$

We omit the details for (68) in this paper.

Define $\widehat{\Sigma}$ by

$$\widehat{\Sigma} = \int_0^1 \widehat{G}_Y(t)\lambda_0(t)dt, \tag{69}$$

where

$$\widehat{G}_Y(t) = \frac{1}{n}\sum_{i=1}^n Y_i(t)\{\boldsymbol{W}_i(t) - \overline{\boldsymbol{W}}_Y(t)\}^{\otimes 2},$$

$$\overline{\boldsymbol{W}}_Y(t) = \frac{n^{-1}\sum_{i=1}^n Y_i(t)\boldsymbol{W}_i(t)}{n^{-1}\sum_{i=1}^n Y_i(t)}.$$

Then by just following the arguments on pp.1161-1162 of [17] with a sufficiently small M, we obtain

$$\widetilde{\Sigma} - \exp\{-C_X C_g\}\{1 + O(L^{-2})\}\widehat{\Sigma} \text{ is n.n.d.} \tag{70}$$

Finally we recall the definitions of $\overline{\Sigma}$, $\overline{G}_Y(t)$, and $\boldsymbol{\mu}_Y(t)$ in Proposition 3 and set

$$\Delta_3 = \widehat{\Sigma} - \overline{\Sigma} = -\int_0^1 \overline{Y}(t)\{\overline{\boldsymbol{W}}_Y(t) - \boldsymbol{\mu}_Y(t)\}^{\otimes 2}\lambda_0(t)dt \tag{71}$$

and $\Delta_4 = \overline{\Sigma} - \mathrm{E}\{\overline{\Sigma}\}$. Then we evaluate

$$\max_{j,k} |(\Delta_3)_{jk}| \quad \text{and} \quad \max_{j,k} |(\Delta_4)_{jk}|.$$

Now we give the details for Δ_1, Δ_3, and Δ_4.

$\boldsymbol{\Delta_1}$: We denote the $(jL + r, kL + m)$ element of $V_n(t, \gamma^*)$ by $v_{jL+r,kL+m}(t)$. Then we have

$$v_{jL+r,kL+m}(t) = (S^{(2)}(t, \gamma^*))_{jL+r,kL+m} - \frac{(S^{(1)}(t, \gamma^*))_{jL+r}(S^{(1)}(t, \gamma^*))_{kL+m}}{S^{(0)}(t, \gamma^*)} \tag{72}$$

and it is easy to see that $|v_{jL+r,kL+m}(t)|$ is uniformly bounded in j, k, r, m, and t. Besides,

$$(S^{(2)}(t, \gamma^*))_{jL+r,kL+m} \leq C_2 \begin{cases} L^{-1}, & r = m = 1 \\ L^{-1/2}|b_r(t)|, & r \geq 2,\ m = 1 \\ L^{-1/2}|b_m(t)|, & r = 1,\ m \geq 2 \\ |b_r(t)||b_m(t)|, & r \geq 2,\ m \geq 2 \end{cases} \tag{73}$$

and

$$(S^{(1)}(t, \gamma^*))_{jL+r} \leq C_3 \begin{cases} L^{-1/2}, & r = 1 \\ |b_r(t)|, & r \geq 2 \end{cases}. \tag{74}$$

By (72)-(74) and some calculation, we evaluate the predictable variation process of Δ_1 and obtain

$$\int_0^1 |v_{jL+r,kL+m}(t)|^2 d < \overline{M}, \overline{M} > (t) \leq \frac{C_4}{n}\int_0^1 |v_{jL+r,kL+m}(t)|\lambda_0(t)dt \leq \frac{C_5}{nL}, \tag{75}$$

where $< \overline{M}, \overline{M} > (t)$ is the predictable variation process of $\overline{M}(t)$. We used (63) here.

Thus we have from the exponential inequality for martingales that

$$\Pr\Big(\max_{j,k} |(\Delta_1)_{jk}| \geq \frac{x(\ln n)^{1/2}}{\sqrt{nL}}\Big) \leq 2(pL)^2 \exp\Big\{-\frac{C_6 x^2 \ln n}{x(\ln n)^{1/2}(n^{-1}L)^{1/2} + 1}\Big\}. \tag{76}$$

$\boldsymbol{\Delta_3}$: Notice that $\overline{\Sigma} - \widehat{\Sigma}$ is n.n.d. Therefore instead of Δ_3, we treat

$$\begin{aligned}\Delta_3' &= \frac{1}{C_Y}\int_0^1 \{\overline{Y}(t)\}^2\{\overline{\boldsymbol{W}}_Y(t) - \boldsymbol{\mu}_Y(t)\}^{\otimes 2}\lambda_0(t)dt \\ &= \frac{1}{C_Y}\int_0^1 \Big[n^{-1}\sum_{i=1}^n \Big\{\boldsymbol{W}_i(t) - Y_i(t)\boldsymbol{\mu}_Y(t)\Big\}\Big]^{\otimes 2}\lambda_0(t)dt.\end{aligned}$$

We evaluate $(\Delta_3')_{kr} = (C_Y n^2)^{-1}\sum_{i,j} f_{ij}$, where $\boldsymbol{\mu}_Y(t) = (\mu_{Y1}(t), \dots, \mu_{Yp}(t))^T$ and

$$f_{ij} = \int_0^1 \{W_{ik}(t) - Y_i(t)\mu_{Yk}(t)\}\{W_{jr}(t) - Y_j(t)\mu_{Yr}(t)\}\lambda_0(t)dt.$$

Note that $|f_{ij}| \le C_7 L^{-1}$. Thus by applying Lemma 4.2 in [17], we obtain

$$\Pr\Big(\max_{k,r}|(\Delta_3')_{kr}| \ge \frac{x(\ln n)^{1/2}}{\sqrt{n}L}\Big) \le 5(pL)^2\exp\Big\{-\frac{C_8 x(n\ln n)^{1/2}}{x^{1/2}(n^{-1}\ln n)^{1/4} + 1}\Big\}. \quad (77)$$

$\boldsymbol{\Delta_4}$: Note that

$$(\overline{\Sigma})_{kr} = \frac{1}{n}\sum_{i=1}^n \int_0^1 Y_i(t)\{W_{ik}(t) - \mu_{Yk}(t)\}\{W_{ir}(t) - \mu_{Yr}(t)\}\lambda_0(t)dt \quad \text{and}$$

$$\Big|\int_0^1 Y_i(t)\{W_{ik}(t) - \mu_{Yk}(t)\}\{W_{ir}(t) - \mu_{Yr}(t)\}\lambda_0(t)dt\Big| \le C_9 L^{-1}.$$

Applying Bernstein's inequality to $(\overline{\Sigma})_{kr}$, we have

$$\Pr\Big(|(\Delta_4)_{kr}| \ge \frac{x(\ln n)^{1/2}}{\sqrt{n}L}\Big) \le 2\exp\Big\{-\frac{C_{10}x^2\ln n}{x(n^{-1}\ln n)^{1/2} + 1}\Big\}.$$

Consequently we have

$$\Pr\Big(\max_{k,r}|(\Delta_4)_{kr}| \ge \frac{x(\ln n)^{1/2}}{\sqrt{n}L}\Big) \le 2(pL)^2\exp\Big\{-\frac{C_{10}x^2\ln n}{x(n^{-1}\ln n)^{1/2} + 1}\Big\}. \quad (78)$$

By combining (67), (68), (70), (71) and (76)-(78) and exploiting Lemma 3, we obtain the desired results. Hence the proof of the proposition is complete.

Finally we verify Corollary 1.

Proof of Corollary 1. Checking Proposition 3 and P_A, P_B, and P_C there, we find we need $x/\{n^{1/5}(\ln n)^{-1/2}\} \to 0$ and $n^{c_p} \sim x^2 \ln n$ to have

$$\kappa^2(\zeta, \ddot{\ell}_p(\gamma^*)) \geq \frac{C_1}{L} + o_p(L^{-1}) \quad \text{and} \quad RE^2(\zeta, \ddot{\ell}_p(\gamma^*)) \geq \frac{C_2}{L} + o_p(L^{-1})$$

for some positive constants C_1 and C_2.

Proposition 2 implies that with probability tending to 1,

$$P_\infty(\ddot{\ell}_p(\gamma^*)) \leq \frac{a_1}{L^{5/2}} + C_3\sqrt{\frac{\ln p}{n}}$$

with $x = C_3(\ln p/\ln n)^{1/2}$ for some positive large constant C_3.

Then we obtain

$$\tau^* \text{ in (30)} \sim L(\ln p/n)^{1/2} \to 0$$

and we have $\eta^* \sim \tau^* \sim L(\ln p/n)^{1/2}$.

Hence the proof of the corollary is complete.

7. Concluding remarks

We proposed an orthonormal basis approach for simultaneous variable selection and structure identification for varying coefficient Cox models. We have derived an oracle inequality for the group Lasso procedure and our method and theory also apply to additive Cox models. These models are among important structured nonparametric regression models. This orthonormal basis approach can be used for the adaptive group Lasso and SCAD. Our simulation study implies that this orthonormal basis approach performs well and that tuning parameter selection by the AIC minimization also works well.

Acknowledgments

We appreciate valuable comments by the associate editor and two reviewers very much. This research is supported by JSPS KAKENHI Grant Number JP 16K05268.

Appendix A. Construction and properties of basis functions

We describe how to construct $\overline{\boldsymbol{B}}(t)$, the properties of $\overline{\boldsymbol{B}}(t)$, and the approximations to $\boldsymbol{g}(t)$. Set

$$\Omega_0 = \int_0^1 \boldsymbol{B}_0(t)\boldsymbol{B}_0^T(t)dt \quad \text{and} \quad \overline{\Omega} = \int_0^1 \overline{\boldsymbol{B}}(t)\overline{\boldsymbol{B}}^T(t)dt.$$

First we describe how to construct A_0 and $\overline{\boldsymbol{B}}(t)$. Set

$$b_1(t) = 1/\sqrt{L} \quad \text{and} \quad b_2(t) = \sqrt{12L^{-1}}(t - 1/2)$$

and define a inner product on the L_2 function space on $[0, 1]$ by

$$(g_1, g_2) = \int_0^1 g_1(t)g_2(t)dt.$$

Then we have

$$\|b_1\|^2 = \|b_2\|^2 = L^{-1} \quad \text{and} \quad (b_1, b_2) = 0.$$

Note that there is some L-dimensional vector $\boldsymbol{a}_{02}$ satisfying $b_2(t) = \boldsymbol{a}_{02}^T\boldsymbol{B}_0(t)$.

We can obtain b_j, $j = 3, \ldots, L$, by just applying the Gram-Schmidt orthonormalization to $(L - 2)$ elements of $\boldsymbol{B}_0(t)$ with the normalization of $\|b_j\|^2 = L^{-1}$. Since every $b_j(t)$ is a linear combination of $\boldsymbol{B}_0(t)$, we have

$$\overline{\boldsymbol{B}}(t) = A_0\boldsymbol{B}_0(t).$$

Hence we have

$$\overline{\Omega} = A_0\Omega_0A_0^T = \begin{pmatrix} 1/L & \boldsymbol{0}^T \\ \boldsymbol{0} & \int \boldsymbol{B}(t)\boldsymbol{B}^T(t)dt \end{pmatrix} = \begin{pmatrix} 1/L & \boldsymbol{0}^T \\ \boldsymbol{0} & A_{-1}\Omega_0A_{-1}^T \end{pmatrix} = \frac{1}{L}I. \tag{A.1}$$

It is known that for some positive constants C_1 and C_2, we have

$$\frac{C_1}{L} \leq \lambda_{\min}(\Omega_0) \leq \lambda_{\max}(\Omega_0) \leq \frac{C_2}{L} \tag{A.2}$$

See Huang et al.[19] for more details.

Thus (A.1) and (A.2) imply that

$$C_3 \leq \lambda_{\min}(A_0A_0^T) \leq \lambda_{\max}(A_0A_0^T) \leq C_4 \tag{A.3}$$

and

$$C_5 \leq \lambda_{\min}(A_{-1}A_{-1}^T) \leq \lambda_{\max}(A_{-1}A_{-1}^T) \leq C_6 \tag{A.4}$$

for some positive constants C_3, C_4, C_5, and C_6. Note that (A.3) implies that

$$C_3 \leq \lambda_{\min}(A_0^T A_0) \leq \lambda_{\max}(A_0^T A_0) \leq C_4.$$

On the other hand, the definition of $\boldsymbol{B}_0(t)$, (A.1), and (A.4) imply that

$$\int_0^1 b_j(t)dt = 0, \text{ for } j = 2, \dots, L, \quad \text{and} \quad \sup_{2 \leq j \leq L} \|b_j\|_\infty = O(1). \tag{A.5}$$

Besides, we have for $\boldsymbol{\gamma}_j = (\gamma_{1j}, \boldsymbol{\gamma}_{-1j}^T)^T \in R^L$,

$$\begin{aligned} \boldsymbol{\gamma}_j^T \overline{\boldsymbol{B}}(t) &= \boldsymbol{\gamma}_j^T A_0 \boldsymbol{B}_0(t) \quad \text{and} \\ |\boldsymbol{\gamma}_j^T \overline{\boldsymbol{B}}(t)| &\leq (\boldsymbol{\gamma}_j^T A_0 A_0^T \boldsymbol{\gamma}_j)^{1/2} |\boldsymbol{B}_0(t)| \leq C_7 |\boldsymbol{\gamma}_j| \end{aligned} \tag{A.6}$$

uniformly on $[0, 1]$ for some positive constant C_7. Note that we used (A.3) and the local property of $\boldsymbol{B}_0(t)$ to derive (A.6).

Next we consider the approximations to $\boldsymbol{g}(t)$. From Corollary 6.26 in [26] and Assumption G, there exist $\boldsymbol{\gamma}_{0j}^* \in R^L$, $j = 1, \dots, p$, satisfying

$$\sum_{j=1}^p \|g_j - \boldsymbol{B}_0^T \boldsymbol{\gamma}_{0j}^*\|_\infty \leq \frac{C_{approx}}{2L^2}, \tag{A.7}$$

where C_{approx} depends on C_g.

In this paper, we use $\overline{\boldsymbol{B}}(t)$ instead of $\boldsymbol{B}_0(t)$. Then

$$\begin{aligned} \boldsymbol{B}_0^T(t)\boldsymbol{\gamma}_{0j}^* &= \overline{\boldsymbol{B}}^T(t)(A_0^T)^{-1}\boldsymbol{\gamma}_{0j}^* = \overline{\boldsymbol{B}}^T(t)\overline{\boldsymbol{\gamma}}_j^* \\ &= \overline{\boldsymbol{B}}^T(t)\begin{pmatrix} \overline{\gamma}_{1j}^* \\ \overline{\boldsymbol{\gamma}}_{-1j}^* \end{pmatrix}, \end{aligned}$$

where $\overline{\boldsymbol{\gamma}}_j^*$, $\overline{\gamma}_{1j}^*$, and $\overline{\boldsymbol{\gamma}}_{-1j}^*$ are defined in the above equations.

Noticing

$$\begin{aligned} &\sum_{j=1}^p \Big| \int_0^1 g_j(t)dt - \frac{\overline{\gamma}_{1j}^*}{L^{1/2}} - \int_0^1 \overline{\boldsymbol{\gamma}}_{-1j}^{*T} \boldsymbol{B}(t)dt \Big| \\ = &\sum_{j=1}^p |g_{cj} - L^{-1/2}\overline{\gamma}_{1j}^*| \leq \frac{C_{approx}}{2L^2}, \end{aligned}$$

we take $\gamma_j^* = 0$ for $\overline{\mathcal{S}}_c \cap \overline{\mathcal{S}}_n$,

$$\begin{aligned}\gamma_{1j}^* = L^{1/2} g_{cj} \quad \text{and} \quad \gamma_{-1j}^* = 0 \quad \text{for } j \in \mathcal{S}_c \cap \overline{\mathcal{S}}_n, \\ \gamma_{1j}^* = L^{1/2} g_{cj} \quad \text{and} \quad \gamma_{-1j}^* = \overline{\gamma}_{-1j}^* \quad \text{for } j \in \mathcal{S}_n.\end{aligned} \tag{A.8}$$

Then from (A.7), we have

$$\sum_{j=1}^{p} \|g_j - \overline{\boldsymbol{B}}^T \gamma_j^*\|_\infty \le \frac{C_{approx}}{L^2} \tag{A.9}$$

and uniformly in j,

$$\begin{aligned}\|g_j\|^2 = |g_{cj}|^2 + \|g_{nj}\|^2 &= \gamma_j^{*T} \overline{\Omega} \gamma_j^* + O(L^{-4}) \\ &= \frac{|\gamma_{1j}^*|^2}{L} + \boldsymbol{\gamma}_{-1j}^{*T} \int_0^1 \boldsymbol{B}(t) \boldsymbol{B}^T(t) dt \boldsymbol{\gamma}_{-1j}^* + O(L^{-4}) \\ &= \frac{|\gamma_{1j}^*|^2}{L} + \frac{|\boldsymbol{\gamma}_{-1j}^*|^2}{L} + O(L^{-4}).\end{aligned}$$

We also have

$$|g_{cj}|^2 = \frac{|\gamma_{1j}^*|^2}{L} \quad \text{and} \quad \|g_{nj}\|^2 = \frac{|\boldsymbol{\gamma}_{-1j}^*|^2}{L} + O(L^{-4}). \tag{A.10}$$

Appendix B. Proofs of technical lemmas

Proof of Lemma 1. From the definitions of $\overline{\boldsymbol{B}}(t)$ and $\boldsymbol{W}_i(t)$. we have

$$\boldsymbol{\theta}^T(\boldsymbol{W}_i(t) - \boldsymbol{W}_j(t)) = \boldsymbol{\theta}^T(I_p \otimes A_0)(\boldsymbol{X}_i(t) \otimes \boldsymbol{B}_0(t) - \boldsymbol{X}_j(t) \otimes \boldsymbol{B}_0(t)). \tag{B.1}$$

Notice that for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T$,

$$|\boldsymbol{\theta}_k^T A_0 \boldsymbol{B}_0(t)| \le |A_0^T \boldsymbol{\theta}_k| \le \{\lambda_{\max}(A_0 A_0^T)\}^{1/2} |\boldsymbol{\theta}_k|. \tag{B.2}$$

Here we used that $|\boldsymbol{B}_0(t)| \le 1$.

Consequently (B.1) and (B.2) yield that

$$\begin{aligned}&|\boldsymbol{\theta}^T(\boldsymbol{W}_i(t) - \boldsymbol{W}_j(t))| \\ &\le \sum_{k=1}^{p} |X_{ik}(t) - X_{jk}(t)| |\boldsymbol{\theta}_k^T A_0 \boldsymbol{B}_0(t)| \\ &\le 2C_X \{\lambda_{\max}(A_0 A_0^T)\}^{1/2} \sum_{k=1}^{p} |\boldsymbol{\theta}_k| \le C_W P_1(\boldsymbol{\theta}).\end{aligned}$$

Hence the proof is complete.

Proof of Lemma 2. This lemma is just a version of Lemma 3.2 in [17]. We can verify this lemma in the same way by taking

$$a_i(t) = \boldsymbol{\theta}^T\{\boldsymbol{W}_i(t) - \widetilde{\boldsymbol{W}}_n(t, \gamma^*)\} \quad \text{and} \quad w_i(t) = Y_i(t)\exp\{\gamma^{*T}\boldsymbol{W}_i(t)\}$$

in the proof. The details are omitted. Hence the proof is complete.

Proof of Lemma 3. This is almost proved in [17]. We should just note that

$$|\gamma^T(\Sigma_1 - \Sigma_2)\gamma| \leq |\gamma|_1^2 \max_{j,k}|(\Sigma_1 - \Sigma_2)_{jk}| \leq L\{P_1(\gamma)\}^2 \max_{j,k}|(\Sigma_1 - \Sigma_2)_{jk}|,$$
$$P_1(\gamma) \leq (1+\zeta)P_1(\gamma_S), \quad \text{and} \quad P_1(\gamma_S) \leq s_0^{1/2}|\gamma|.$$

When $\Sigma_2 - \Sigma_1$ is n.n.d., we have

$$|\gamma^T(\Sigma_1 - \Sigma_2)\gamma| \leq \gamma^T\Delta\gamma \leq L\{P_1(\gamma)\}^2 \max_{j,k}|(\Delta)_{jk}|.$$

Hence the proof is complete.

Appendix C. Derivatives of the B-spline basis

In this section, we examine properties of

$$\int_0^1 \boldsymbol{B}_0'(t)(\boldsymbol{B}_0'(t))^T dt \tag{C.1}$$

and describe why we have adopted the orthogonal decomposition approach while the other authors have considered the L_2 norm of the estimated derivatives when they deal with structure identification for additive models or partially linear additive models.

We take a function $g_A(t)$ on $[0, 1]$ defined by

$$g_A(t) = \sin(2\pi At)$$

for $A \to \infty$ sufficiently slowly. Then it is easy to see

$$\|g_A\|^2 \sim 1, \quad \|g_A'\|^2 \sim A^2, \quad \text{and} \quad \|g_A''\|^2 \sim A^4.$$

On the other hand, we can approximate this $g_A(t)$ by $\boldsymbol{B}_0(t)\gamma_A$ accurately enough and we have

$$\gamma_A^T\Omega_0\gamma_A \sim 1, \quad |\gamma_A|^2 \sim L, \quad \text{and} \quad \gamma_A^T\int_0^1 \boldsymbol{B}_0'(t)(\boldsymbol{B}_0'(t))^T dt\gamma_A \sim A^2 \to \infty.$$

This means some eigenvalues of the matrix defined in (C.1) have the order larger than L^{-1}.

Hence we cannot follow the proofs given in those papers based on the L_2 norm of the estimated derivatives. This is because the eigenvalue property just proved in this paper violates their assumptions on matrices similar to

$$\int_0^1 \boldsymbol{B}_0''(t)(\boldsymbol{B}_0''(t))^T dt.$$

The above matrix also should have some larger eigenvalues as that in (C.1). Besides, it is more difficult to estimate the derivatives of the coefficient functions. This is why we have adopted the orthogonal decomposition approach. Zhang et al.[35] is based on the smoothing spline method and it is difficult to apply their ingenious approach to the loss function other than the L_2 loss function.

Appendix D. Proofs for other models

We outline necessary changes in the proofs for the former model in section 4 since both models in the section can be treated in almost the same way as the time-varying coefficient model. Especially, almost no change is necessary to the proofs of Proposition 1 and Theorem 1.

We assume standard assumptions for varying coefficient models here.

Proof of Proposition 2) The poof consists of (56)-(60) and (61)-(65).

(56)-(60): Note that $|b_{0j}(t)|$ is replaced with $n^{-1}\sum_{i=1}^n |b_{0j}(Z_i(t))|$. When we evaluate the predicable variation process in (58),

$$\int_0^1 |b_{0j}(t)|^2 \lambda_0(t)dt \le C\int_0^1 |b_{0j}(t)|\lambda_0(t)dt$$

is replaced with

$$\int_0^1 \Big\{n^{-1}\sum_{i=1}^n |b_{0j}(Z_i(t))|\Big\}^2 \lambda_0(t)dt \le C\int_0^1 n^{-1}\sum_{i=1}^n b_{0j}^2(Z_i(t))\lambda_0(t)dt. \qquad \text{(D.1)}$$

We can evaluate the second term in (D.1) by using Bernstein's inequality and

$$\mathrm{E}\Big\{n^{-1}\int_0^1 \sum_{i=1}^n b_{0j}^2(Z_i(t))\lambda_0(t)dt\Big\} = \int_0^1 \mathrm{E}\{b_{0j}^2(Z_1(t))\}\lambda_0(t)dt = O(L^{-1}).$$

(61)-(65): When we apply the martingale exponential inequality, (63) is replaced with

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 b_j^2(Z_i(t))\lambda_0(t)dt.$$

We can evaluate this expression by using Bernstein's inequality and

$$\begin{aligned}\mathrm{E}\Big\{\int_0^1 b_j^2(Z_1(t))\lambda_0(t)dt\Big\} &\le C\boldsymbol{a}_{0j}^T\int_0^1 \mathrm{E}\{\boldsymbol{B}_0(Z_1(t))(\boldsymbol{B}_0(Z_1(t)))^T\}\lambda_0(t)dt\boldsymbol{a}_{0j}\\ &= O(L^{-1}).\end{aligned}$$

We need some assumptions for $\mathrm{E}\{\boldsymbol{B}_0(Z_1(t))(\boldsymbol{B}_0(Z_1(t)))^T\}$ as for Ω_0 in Appendix A.

Proof of Proposition 3) The proof consists of evaluating Δ_1, Δ_3, and Δ_4.

$\boldsymbol{\Delta_1}$: We should just follow the line of (61)-(65).

$\boldsymbol{\Delta_3}$: This is almost a U-statistic and we can also apply the exponential inequality for U-statistics as (3.5) in [12] to the part of a U-statistic.

$\boldsymbol{\Delta_4}$: This is a sum of bounded independent random variables and we can deal with this by applying Bernstein's inequality.

References

[1] Belloni, A. and V. Chernozhukov, V. *l*1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics 39(2011) 82-130.

[2] Breheny, P. The R package 'grpreg' : Regularization Paths for Regression Models with Grouped Covariates. Version 3.0-2 (2016).

[3] Bickel, P. J., Ritov, Y. A., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37(2009) 1705-1732.

[4] Bradic, J., Fan, J., and Jiang, J. Regularization for Cox's proportional hazards model with NP-dimensionality. The Annals of Statistics 39(2011) 3092-3120.

[5] Bradic, J. and Song, R. Structured estimation for the nonparametric Cox model. Electronic Journal of Statistics 9(2015) 492-534.

[6] Bühlmann, P. and van de Geer, S. Statistics for High-dimensional Data: Methods, Theory and Applications. Springer Science & Business Media. 2011.

[7] Cai, J., Fan, J., Zhou, H., and Zhou, Y. Hazard models with varying coefficients for multivariate failure time data. The Annals of Statistics 35(2007) 324-354.

[8] Cai, Z. and Sun, Y. Local linear estimation for time-dependent coefficients in Cox's regression models. Scandinavian Journal of Statistics 30(2003) 93-111.

[9] Cox, D. R. Regression models and life tables (with discussion). Journal of the Royal Statistical Society Series B 34(1972) 187-220.

[10] Fan, J., Fan, Y., and Barut, E. Adaptive robust variable selection. The Annals of Statistics 42(2014) 324-351.

[11] Fan, J., Xue, L., and Zou, H. Strong oracle optimality of folded concave penalized estimation. The Annals of Statistics 42(2014) 819-849.

[12] Giné, E., Latała, R., and Zinn, J. Exponential and moment inequalities for U-statistics. In High Dimensional Probability II (pp. 13-38). Birkhäuser. 2000.

[13] Guilloux, A., Lemler, S., and Taupin, M.L. Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates. Journal of Multivariate Analysis, 148(2016) 141-159.

[14] Hastie, T., Tibshirani, R., and Wainwright, M. Statistical Learning with Sparsity: the Lasso and Generalizations. CRC Press. 2015.

[15] Honda, T. and Härdle, W. K. Variable selection in Cox regression models with varying coefficients. Journal of Statistical Planning and Inference 148(2014) 67-81.

[16] Huang J., Breheny, P., and Ma, S. A selective review of group selection in high dimensional models. Statistical Science 27(2012) 481-499

[17] Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C. H. Oracle inequalities for the lasso in the Cox model. The Annals of Statistics 41(2013) 1142-1165.

[18] Huang, J. Z., Kooperberg, C., Stone, C. J., and Truong, Y. K. Functional ANOVA modeling for proportional hazards regression. The Annals of Statistics 28(2000) 961-999.

[19] Huang, J. Z., Wu, C. O., and Zhou, L. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statistitica Sinica 14(2004) 763-788.

[20] Kalbfleisch, J. D. and Prentice, R. L. The Statistical Analysis of Failure Time Data, Second Edition. Wiley. 2002.

[21] Kato, K. Group Lasso for High Dimensional Sparse Quantile Regression Models. arXiv preprint arXiv:1103.1458. 2011

[22] Kong, S. and Nan, B. Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. Statistica Sinica 24 (2014) 25-42

[23] Lemler, S. Oracle inequalities for the Lasso in the high-dimensional Aalen multiplicative intensity model. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 52(2016) 981-1008.

[24] Lian, H., Lai, P., and Liang, H. Partially linear structure selection in Cox models with varying coefficients. Biometrics 69(2013) 348-357.

[25] Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. Oracle inequalities and optimal inference under group sparsity. The Annals of Statistics 39(2011) 2164-2204.

[26] Schumaker, L. Spline Functions: Basic Theory, Third Edition. Cambridge University Press. 2007

[27] Song, R., Lu, W., Ma, S., and Jeng, X. J. Censored rank independence screening for high-dimensional survival data. Biometrika 101(2014) 799-814.

[28] Sun, H., Lin, W., Feng, R. and Li, H. Network-regularized high-dimensional cox regression for analysis of genomic data. Statistica Sinica 24(2014) 1433-1459.

[29] Tang, Y., Song, X., Wang, H. J., and Zhu, Z. Variable selection in high-dimensional quantile varying coefficient models. Journal of Multivariate Analysis 122(2013) 115-132.

[30] van de Geer, S. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. The Annals of Statistics 23(1995) 1779-1801.

[31] van der Vaart, A. W. and Wellner, J. A. Weak Convergence and Empirical Processes. Springer. 1996.

[32] Wang, S., Nan, B., Zhu, N., and Zhu, J. Hierarchically penalized Cox regression with grouped variables. Biometrika 96(2009) 307-322.

[33] Yan, J. and Huang, J. Model selection for Cox models with time-varying coefficients. Biometrics 68(2012) 419-428.

[34] Yang, G., Yu, Y., Li, R., and Buu, A. Feature screening in ultrahigh dimensional Cox's model. Forthcoming in Statistica Sinica.

[35] Zhang, H. H., Cheng, G., and Liu, Y. Linear or nonlinear? Automatic structure discovery for partially linear models. Journal of the American Statistical Association. 106(2012) 1099-1112.

[36] Zhang, H. H. and Lu, W. Adaptive Lasso for Cox's proportional hazards model. Biometrika 94(2007) 691-703.

[37] Zhang, S., Wang, L., and Lian, H. Estimation by polynomial splines with variable selection in additive Cox models. Statistics 48(2014) 67-80.

[38] Zhao, J. and Leng, C. An analysis of penalized interaction models. Bernoulli 22(2016) 1937-1961.

[39] Zhao, P., Rocha, G., and Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics 37(2009) 3468-3497.

[40] Zhao, S. D. and Li, Y. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. Journal of Multivariate Analysis 105(2012) 397-411.

[41] Zou, H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101(2006) 1418-1429.

Supplement to "Variable selection and structure identification for varying coefficient Cox models"

by Toshio Honda and Ryota Yabe

Appendix E. Hierarchical penalty

We give an expression of $\nabla_j P_h(\gamma)$. Recall that

$$\nabla_j P_h(\gamma) = \nabla_j(|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{1/q} + \nabla_j|\gamma_{-1j}|.$$

Set

$$\nabla_j(|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{1/q} = \begin{pmatrix} d_{1j} \\ \boldsymbol{d}_{-1j} \end{pmatrix},$$

where $d_{1j} \in R$ and $\boldsymbol{d}_{-1j} \in R^{L-1}$.

When $|\gamma_{1j}| = 0$ and $|\gamma_{-1j}| = 0$,

$$d_{1j} = \epsilon_{1j} \quad \text{and} \quad \boldsymbol{d}_{-1j} = \boldsymbol{\epsilon}_{-1j},$$

where $|\epsilon_{1j}| \leq a$ and $|\boldsymbol{\epsilon}_{-1j}| \leq b$ such that (a, b) satisfies $(1 + t^q)^{1/q} \geq a + bt$ for any $t \geq 0$. This follows from the definition of subgradient and we note that $0 \leq a \leq 1$ and $0 \leq b \leq 1$.

When $|\gamma_{1j}| \neq 0$ and $|\gamma_{-1j}| = 0$,

$$d_{1j} = \text{sign}(\gamma_{1j}) \quad \text{and} \quad \boldsymbol{d}_{-1j} = 0.$$

When $|\gamma_{1j}| = 0$ and $|\gamma_{-1j}| \neq 0$,

$$d_{1j} = 0 \quad \text{and} \quad \boldsymbol{d}_{-1j} = \gamma_{-1j}/|\gamma_{-1j}|. \tag{E.1}$$

This property is essential to hierarchical selection for g_{cj} and $g_{nj}(t)$. See [39].

When $|\gamma_{1j}| \neq 0$ and $|\gamma_{-1j}| \neq 0$,

$$d_{1j} = (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{\frac{1}{q}-1}\text{sign}(\gamma_{1j})|\gamma_{1j}|^{q-1}$$

and

$$\boldsymbol{d}_{-1j} = (|\gamma_{1j}|^q + |\gamma_{-1j}|^q)^{\frac{1}{q}-1}\frac{\gamma_{-1j}}{|\gamma_{-1j}|}|\gamma_{-1j}|^{q-1}.$$

We state a version of Proposition 1 for $Q_h(\gamma; \lambda)$. We state this proposition, Proposition 4, in terms of $P_1(\gamma)$. This is essential in proving the oracle inequality for $Q_h(\gamma; \lambda)$ and they are not any typos. Once this proposition is established, we can proceed exactly in the same way as for $Q_1(\gamma; \lambda)$ with changes of some constants.

Proposition 4. *If $\lambda > D_\ell$, we have*

$$(\widehat{\gamma} - \gamma^*)^T\{\dot{l}_p(\widehat{\gamma}) - \dot{l}_p(\gamma^*)\} \le (2\lambda + D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}) - (\lambda - D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}})$$

and

$$(\lambda - D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}}) \le (2\lambda + D_\ell)P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}).$$

Therefore if $D_\ell \le \xi\lambda\,(\xi < 1)$, we have

$$P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}}) \le \frac{2+\xi}{1-\xi}P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}).$$

Proof) Note that

$$\begin{aligned}
&(\widehat{\gamma} - \gamma^*)^T(\dot{l}_p(\widehat{\gamma}) - \dot{l}_p(\gamma^*)) \qquad\qquad (E.2)\\
&= \Big\{\sum_{j\in\overline{\mathcal{S}}_c}\widehat{\theta}_{1j}\frac{\partial\ell_p}{\partial\gamma_{1j}}(\widehat{\gamma}) + \sum_{j\in\overline{\mathcal{S}}_c}\widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial\ell_p}{\partial\boldsymbol{\gamma}_{-1j}}(\widehat{\gamma})\Big\}\\
&\quad+\Big\{\sum_{j\in\overline{\mathcal{S}}_n\cap\mathcal{S}_c}\widehat{\theta}_{1j}\frac{\partial\ell_p}{\partial\gamma_{1j}}(\widehat{\gamma}) + \sum_{j\in\overline{\mathcal{S}}_n\cap\mathcal{S}_c}\widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial\ell_p}{\partial\boldsymbol{\gamma}_{-1j}}(\widehat{\gamma})\Big\}\\
&\quad+\Big\{\sum_{j\in\mathcal{S}_n}\widehat{\theta}_{1j}\frac{\partial\ell_p}{\partial\gamma_{1j}}(\widehat{\gamma}) + \sum_{j\in\mathcal{S}_n}\widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial\ell_p}{\partial\boldsymbol{\gamma}_{-1j}}(\widehat{\gamma})\Big\}\\
&\quad+\{-\widehat{\boldsymbol{\theta}}^T(\dot{l}_p(\gamma^*)\} = E_1 + E_2 + E_3 + E_4 \ge 0,
\end{aligned}$$

where E_j, $j = 1, \ldots, 4$, are defined in the above equation.

The last inequality follows from the convexity of $\ell_p(\gamma)$ and we should recall that $\widehat{\boldsymbol{\theta}} = \widehat{\gamma} - \gamma^*$.

We evaluate $E_j,\ j = 1, 2, 3, 4$.

E$_1$: Notice that $\widehat{\gamma}_j = \widehat{\boldsymbol{\theta}}_j$. Then we should evaluate

$$E_{1j} = \widehat{\theta}_{1j}\frac{\partial\ell_p}{\partial\gamma_{1j}}(\widehat{\gamma}) + \widehat{\boldsymbol{\theta}}^T_{-1j}\frac{\partial\ell_p}{\partial\boldsymbol{\gamma}_{-1j}}(\widehat{\gamma}).$$

When $\widehat{\gamma}_{1j} \ne 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} \ne 0$, we have

$$E_{1j} = -\lambda(|\widehat{\theta}_{1j}|^q + |\widehat{\boldsymbol{\theta}}_{-1j}|^q)^{1/q} - \lambda|\widehat{\boldsymbol{\theta}}_{-1j}|. \qquad (E.3)$$

When $\widehat{\gamma}_{1j} \neq 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} = 0$, we have

$$E_{1j} = -\lambda|\widehat{\theta}_{1j}|. \tag{E.4}$$

When $\widehat{\gamma}_{1j} = 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} \neq 0$, we have

$$E_{1j} = -2\lambda|\widehat{\boldsymbol{\theta}}_{-1j}|. \tag{E.5}$$

From (E.3)-(E.5), we obtain

$$E_1 \leq -\lambda \sum_{j \in \overline{\mathcal{S}}_c} (|\widehat{\theta}_{1j}| + |\widehat{\boldsymbol{\theta}}_{-1j}|). \tag{E.6}$$

$\mathbf{E_2}$: First notice that

$$\widehat{\boldsymbol{\gamma}}_{-1j} = \widehat{\boldsymbol{\theta}}_{-1j} \quad \text{and} \quad |\frac{\partial \ell_p}{\partial \gamma_{1j}}(\widehat{\boldsymbol{\gamma}})| \leq \lambda$$

and we should evaluate

$$E_{2j} = \widehat{\theta}_{1j} \frac{\partial \ell_p}{\partial \gamma_{1j}}(\widehat{\boldsymbol{\gamma}}) + \widehat{\boldsymbol{\gamma}}_{-1j}^T \frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\boldsymbol{\gamma}}).$$

When $\widehat{\gamma}_{1j} \neq 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} \neq 0$, we have

$$\begin{aligned} E_{2j} &\leq \lambda|\widehat{\theta}_{1j}| - \lambda(|\widehat{\gamma}_{ij}|^q + |\widehat{\boldsymbol{\theta}}_{-1j}|^q)^{\frac{1}{q}-1}|\widehat{\boldsymbol{\theta}}_{-1j}|^q - \lambda|\widehat{\boldsymbol{\theta}}_{-1j}| \\ &\leq \lambda(|\widehat{\theta}_{1j}| - |\widehat{\boldsymbol{\theta}}_{-1j}|). \end{aligned} \tag{E.7}$$

When $\widehat{\gamma}_{1j} \neq 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} = 0$, we have

$$E_{2j} \leq \lambda|\widehat{\theta}_{1j}|. \tag{E.8}$$

When $\widehat{\gamma}_{1j} = 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} \neq 0$ and when $\widehat{\gamma}_{1j} = 0$ and $\widehat{\boldsymbol{\gamma}}_{-1j} = 0$, we have

$$E_{2j} \leq \lambda|\widehat{\theta}_{1j}| - 2\lambda|\widehat{\boldsymbol{\theta}}_{-1j}|. \tag{E.9}$$

From (E.7)-(E.9), we obtain

$$E_2 \leq \lambda \sum_{j \in \overline{\mathcal{S}}_n \cap \mathcal{S}_c} (|\widehat{\theta}_{1j}| - |\widehat{\boldsymbol{\theta}}_{-1j}|) \leq \lambda \sum_{j \in \overline{\mathcal{S}}_n \cap \mathcal{S}_c} (2|\widehat{\theta}_{1j}| - |\widehat{\boldsymbol{\theta}}_{-1j}|). \tag{E.10}$$

E₃ : Notice that

$$\frac{\partial \ell_p}{\partial \gamma_{1j}}(\widehat{\gamma})| \leq \lambda \quad \text{and} \quad |\frac{\partial \ell_p}{\partial \boldsymbol{\gamma}_{-1j}}(\widehat{\gamma})| \leq 2\lambda.$$

Then we have

$$E_3 \leq 2\lambda \sum_{j \in \mathcal{S}_n} (|\widehat{\theta}_{1j}| + |\widehat{\boldsymbol{\theta}}_{-1j}|). \tag{E.11}$$

E₄ : We have

$$E_4 \leq P_1(\widehat{\boldsymbol{\theta}}) D_\ell = (P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}) + P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}})) D_\ell. \tag{E.12}$$

(E.6), (E.10), (E.11), and (E.12) yield that

$$E_1 + E_2 + E_3 + E_4 \leq (2\lambda + D_\ell) P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}) - (\lambda - D_\ell) P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}}).$$

The first and second inequalities follow from (E.2) and the above inequality. The third inequality follows from the following expression of the second one.

$$P_1(\widehat{\boldsymbol{\theta}}_{\overline{\mathcal{S}}}) \leq \frac{2\lambda + D_\ell}{\lambda - D_\ell} P_1(\widehat{\boldsymbol{\theta}}_{\mathcal{S}})$$

Hence the proof of the proposition is complete.

Appendix F. Additional simulation results

In this appendix, we present the following.
1. BIC minimization results for the simulations in section 5
2. Estimation error results for the simulations in section 5
3. SCAD results for the simulations in section 5
4. Simulation results for another varying coefficient model

BIC results: The results for the group Lasso with the BIC minimization are given in Tables F.9 and F.10. The group Lasso with the BIC minimization does not work well because it tends to remove relevant covariates.

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q(q = 8)$		X_{q+1} to X_p	
$p = 500$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.080	—	0.000	0.790	—	—	—	—
Correct	0.920	1.000	1.000	0.210	1.000	1.000	0.996	1.000
FPR	—	0.000	—	—	0.000	0.000	0.004	0.000
$p = 300$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.062	—	0.004	0.719	—	—	—	—
Correct	0.938	1.000	0.996	0.281	0.988	1.000	0.994	1.000
FPR	—	0.000	—	—	0.012	0.000	0.006	0.000
$p = 150$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.058	—	0.001	0.616	—	—	—	—
Correct	0.942	1.000	0.999	0.384	0.986	1.000	0.989	1.000
FPR	—	0.000	—	—	0.014	0.000	0.011	0.000
$p = 50$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.030	—	0.001	0.332	—	—	—	—
Correct	0.970	0.998	0.999	0.668	0.959	0.998	0.961	0.999
FPR	—	0.002	—	—	0.041	0.002	0.039	0.001

Table F.9: Varying coefficient model with an index variable(BIC)

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q(q = 8)$		X_{q+1} to X_p	
$p = 500$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.195	—	0.695	0.435	—	—	—	—
Correct	0.805	1.000	0.305	0.565	1.000	0.998	1.000	1.000
FPR	—	0.000	—	—	0.000	0.002	0.000	0.000
FNR	0.134	—	0.631	0.345	—	—	—	—
Correct	0.866	0.998	0.369	0.655	1.000	1.000	1.000	0.999
FPR	—	0.002	—	—	0.000	0.000	0.000	0.001
$p = 150$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.080	—	0.499	0.222	—	—	—	—
Correct	0.920	0.995	0.501	0.778	0.999	0.998	0.999	0.997
FPR	—	0.005	—	—	0.001	0.002	0.001	0.003
$p = 50$	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.034	—	0.348	0.105	—	—	—	—
Correct	0.966	0.991	0.652	0.895	0.999	0.991	0.998	0.991
FPR	—	0.009	—	—	0.001	0.009	0.002	0.009

Table F.10: Additive model(BIC)

Estimation error: We show the estimation errors of the AIC minimum and oracle estimators in Figure F.1 for the varying coefficient model and Figure F.2 for the additive model. These figures are the box plots of

$$\sqrt{\sum_{j=1}^{p} \|\widehat{g}_j - g_j\|^2}$$

by the AIC minimization group Lasso (AIC) and the oracle estimator (coxph). Note that we used the coxph function and the knowledge of the true models for the oracle estimator.

As shown in the figures, the group Lasso may not be a good estimator of the parameters or functions since they are biased in spite of its nice theoretical properties. We think we should use the group Lasso as a tool of variable selection or simultaneous variable selection and structure identification because it showed very good performances for these purposed in our numerical studies. We should do some kind of debiasing as in
van de Geer, S., Bühlmann, P., Ritov, Y.A. and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics 42(2014), pp.1166-1202.
However, it is a topic of future research for more complicated models than linear models.

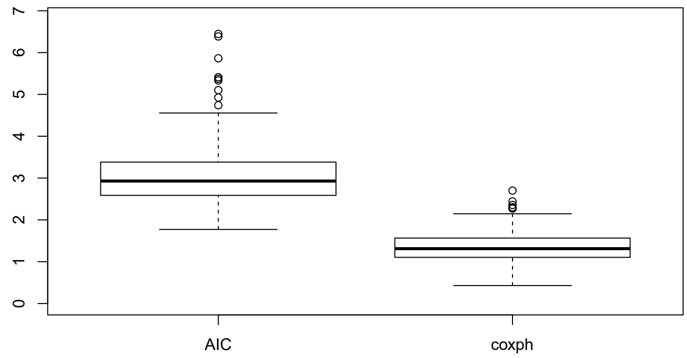

Figure F.1: Estimation error for the varying coefficient model

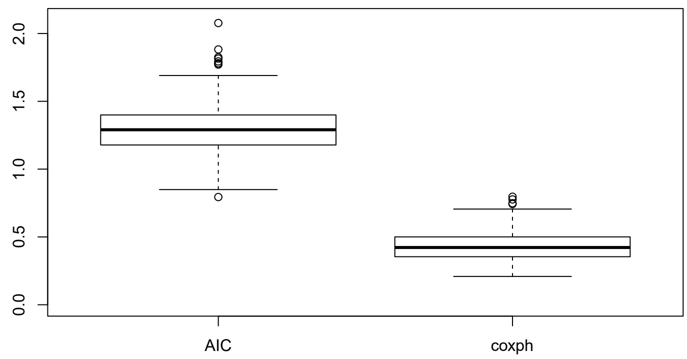

Figure F.2: Estimation error for the additive model

SCAD results: The results for the group SCAD are given in F.11 and F.12. The models are the same ones as in section 5. We took only $p = 50$ since the results for $p = 150$ and $p = 300$ are unstable and very bad. Probably the minimization of the grpsurv function does not work and this is due to the nonconvexity of the SCAD penalty. That is why various kinds of screening procedures have been proposed to give suitable initial values or reduce the numbers of covariates.

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q (q = 8)$		X_{q+1} to X_p	
AIC	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.116	—	0.022	0.118	—	—	—	—
Correct	0.884	0.778	0.978	0.882	0.979	0.789	0.974	0.827
FPR	—	0.222	—	—	0.021	0.211	0.026	0.173
BIC	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.066	—	0.019	0.230	—	—	—	—
Correct	0.934	0.991	0.981	0.770	0.988	0.994	0.992	0.996
FPR	—	0.009	—	—	0.012	0.006	0.008	0.004

Table F.11: Varying coefficient model with an index variable(SCAD, $p = 50$)

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q (q = 8)$		X_{q+1} to X_p	
AIC	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.015	—	0.145	0.002	—	—	—	—
Correct	0.985	0.934	0.855	0.998	0.991	0.922	0.990	0.933
FPR	—	0.066	—	—	0.009	0.078	0.010	0.067
BIC	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear	Linear	Nonlinear
FNR	0.040	—	0.264	0.058	—	—	—	—
Correct	0.960	0.986	0.736	0.942	0.998	0.976	0.996	0.978
FPR	—	0.014	—	—	0.002	0.024	0.004	0.022

Table F.12: Additive model(SCAD, $p = 50$)

Another varying coefficient model: We replaced $g_3(z)$ and $g_4(z)$ of the varying coefficient model in section 5 with

$$g_3(z) = 3\{2^{-1/2}\cos(2\pi z) + (z - 1/2)\} \quad \text{and} \quad g_4(z) = 3\sin(2\pi z),$$

respectively. We didn't change the other setup including the censoring variable. Then the censoring rate is about 45%. We presented the results for our group Lasso procedure with AIC minimization in Table F.13. Both X_3 and X_4 have no constant component. We have a rather high false discovery rate for the constant components for X_3 and X_4. The BIC minimization didn't perform well for this model, either and we omitted the BIC results.

$n = 300$	X_1 and X_2		X_3 and X_4		X_5 to $X_q (q = 8)$		X_{q+1} to X_p	
$p = 500$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.090	—	—	0.070	—	—	—	—
Correct	0.910	0.990	0.855	0.930	0.958	0.998	0.956	0.997
FPR	—	0.010	0.145	—	0.042	0.002	0.044	0.003
$p = 300$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.054	—	—	0.051	—	—	—	—
Correct	0.946	0.984	0.826	0.949	0.948	0.995	0.947	0.994
FPR	—	0.016	0.174	—	0.052	0.005	0.053	0.006
$p = 150$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.038	—	—	0.049	—	—	—	—
Correct	0.962	0.974	0.836	0.951	0.935	0.982	0.931	0.988
FPR	—	0.026	0.164	—	0.065	0.018	0.069	0.012
$p = 50$	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.	Const.	Non-const.
FNR	0.021	—	—	0.014	—	—	—	—
Correct	0.979	0.916	0.795	0.986	0.903	0.956	0.890	0.961
FPR	—	0.084	0.205	—	0.097	0.044	0.110	0.039

Table F.13: Another varying coefficient model with an index variable(AIC)

Appendix G. More details on real data analysis

First we give more details on artificial covariates. Let R_j, $j = 1, 2, \ldots$, independently follow the standard normal distribution in the standardized case and the uniform distribution on $[0, 1]$ in the transformed case, respectively.
X_9 and X_{10} : They follow the Bernoulli distribution with $\Pr(X_j = 1) = 0.5$ independently of each other and all the other variables.
$X_{11}, \ldots, X_{14}$: We define them by $X_{10+j} = \rho X_{4+j} + (1 - \rho)R_j$ with $\rho = 0.2$ for $j = 1, 2, 3, 4$.
$X_{15}, \ldots, X_p$: We define them by $X_{10+j} = R_j$ for $j = 5, \ldots, p$.

Next we examine the design matrix. We define a matrix D by

$$D = n^{-1} X_D^T X_D,$$

where X_D is a $n \times 5$ matrix and its first column consists of 1 and the other columns consist of $X_1, \ldots, X_4$. Its maximum eigenvalue is 2.051 and its minimum one is 0.035. Thus $\lambda_{\max}(D)/\lambda_{\min}(D) = 2.051/0.035$ is larger than 58. Even if we remove $X_1 = tgrad2$, the ratio is still more than 11. This also suggests serious multicollinearity among dummy variables.

Finally we describe the transformation of continuous variables. We examined the histograms and minimum values of continuous variables and then transformed them so that they look uniformly distributed on $[0, 1]$. Specifically,

$$\begin{aligned}
X_5 &= \mathrm{PCHI}(tsize, df = m_{tsize}) \\
X_6 &= \mathrm{PEXP}(pnodes, rate = 0.13) \\
X_7 &= \mathrm{PEXP}(progrec, rate = 1/m_{progrec}) \\
X_8 &= \mathrm{PEXP}(estrec, rate = 1/m_{estrec}),
\end{aligned}$$

where $m_{variable}$ is the mean of the variable, $\mathrm{PCHI}(x, df = m)$ is the distribution function of the chi-squared distribution with $df = m$, and $\mathrm{PEXP}(x, rate = 1/m)$ is the distribution function of the exponential distribution with mean$= m$.

This is the end of the supplement.