SFB 649 Discussion Paper 2012-061

Variable selection in Cox regression models with varying coefficients

Toshio Honda *
Wolfgang Karl Härdle **

* Graduate School of Economics, Hitotsubashi University
** Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

Variable selection in Cox regression models with varying coefficients[☆]

Toshio Honda[*]

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

Wolfgang Karl Härdle

C.A.S.E. - Center for Applied Statistics & Economics, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany

Abstract

We deal with two kinds of Cox regression models with varying coefficients. The coefficients vary with time in one model. In the other model, there is an important random variable called an index variable and the coefficients vary with the variable. In both models, we have p-dimensional covariates and p increases moderately. However, it is the case that only a small part of the covariates are relevant in these situations. We carry out variable selection and estimation of the coefficient functions by using the group SCAD-type estimator and the adaptive group Lasso estimator. We examine the theoretical properties of the estimators, especially the L_2 convergence rate, the sparsity, and the oracle property. Simulation studies and a real data analysis show the performance of these new techniques.

Keywords: Cox regression model, high-dimensional data, sparsity, oracle estimator, B-splines, group SCAD, adaptive group Lasso, L_2 convergence rate
JEL: C14, C24

[☆]This research is supported by the Deutsche Forschungsgemeinschaft via SFB 649 "Economic Risk", Humboldt-Universität zu Berlin.

[*]The first author is also supported by the Global COE Program Research Unit for Statistical and Empirical Analysis in Social Sciences at Hitotsubashi University, Japan.

Email address: honda@econ.hit-u.ac.jp (Toshio Honda)

1. Introduction

The Cox regression model is one of the most popular and useful models in survival analysis. In recent years, many nonparametric and semiparametric variants of the Cox regression model have been proposed. Among them, there are varying coefficient models (Cai and Sun [9], Tian et al. [24], Fan et al. [12], Cai et al. [8], Chen et al. [10]), partially linear models and their extensions (Cai et al. [5], [6]), and additive and functional ANOVA models (Huang et al. [17]). In this paper we focus on varying coefficient models and consider two kinds of Cox regression models with varying coefficients. The coefficients vary with time in one model ([9], [24]) and with index variable $U(t)$ in another model ([12], [8], [10]).

In recent years, a dimensional and model selection issue occurs in many applications : only a small part of the variables are relevant. Therefore statistical methods for variable selection are needed. Penalized likelihood estimators such as the Lasso or SCAD estimators have been among the standard tools in carrying out variable selection and estimation simultaneously, Tibshirani [25] and Fan and Li [13]. Zou [33] proposed the adaptive Lasso to correct some deficiencies of the Lasso and proved that the adaptive Lasso estimators choose the relevant variables consistently. Both group SCAD and group Lasso are also popular techniques, Yuan and Lin [27] and Meier et al. [21]. For more references, we refer to Bühlmann and van de Geer [4].

Local linear estimators have been used in varying coefficient Cox regression models. However, we employ basis functions such as B-spline basis functions and look at the models from a different perspective by combining these models and variable selection. This is the focus of this research. We deal with the cases where the number of the covariates, p, increase moderately with the sample size, for example $p = \mathcal{O}(n^{3/10})$, where n is the sample size. We conduct variable selection and estimation simultaneously by employing group SCAD-type or adaptive group Lasso estimators.

Variable selection and estimation in Cox regression models are considered in many papers, for example, in Cai et al. [7], Zhang and Lin [32], Du et al. [11], Wang et al. [29], and Bradic et al. [2]. In [29], group SCAD and Lasso estimators are analyzed for linear models. Bradic et al. [2] deals with ultra high-dimensional data and presents useful theoretical results for linear models. However, all of the above papers focus on variable selection in the parametric part of the Cox regression model and the variants. Although an alleviation has been proposed for functional ANOVA models in Leng and

Zhang [19], the sparsity or the oracle property of the estimator is not verified.

Recently Yan and Huang [26] proposed the adaptive group Lasso in a Cox regression model with time-varying coefficients and carried out some simulation studies and a real data analysis. There is however still a lacuna of theoretical results that this research aims to fill. We establish the sparsity for the group SCAD-type and adaptive group Lasso estimator and the oracle property for the group SCAD-type estimator under simple and interpretable assumptions. Very recently Bradic and Song [3] considered penalized estimators for additive Cox regression models when the number of the covariates is larger than the sample size. However, they do not deal with any varying coefficient Cox regression models.

We concentrate on the time-varying coefficient model since the model is easier to treat. Besides no identifiability constraint is necessary to the model. We describe only the results on the model having coefficients varying with an index variable $U(t)$ in section 6 because we can deal with the model almost in the same way. The derivation of the theoretical results of this paper crucially depend on the methodology of Huang [16], Huang and Stone [18], and Huang et al. [17].

Variable selection in time-varying coefficients models are also considered in other settings in Wang et al. [28], Noh and Park [22], Wei et al. [30], and Lian [20], where group SCAD-type or group Lasso estimators are used.

This paper is organized as follows. We state the setup of the time-varying coefficient model and define the partial likelihood estimator, the group SCAD-type estimator, and the adaptive group Lasso estimator in section 2. We consider the asymptotics and establish the sparsity and the oracle property of the estimators in section 3. The results of simulation studies and a real example are presented in section 4. Technical assumptions and the proofs of the theorems are given in section 5. We present the results on the model having coefficients varying with an index variable $U(t)$ in section 6. The proofs of propositions and lemmas are confined to section 7.

In this paper, C is a generic positive constant and the value varies from place to place. We denote the Euclidean norm and the transpose of a vector v by $|v|$ and v^T, respectively. We omit almost surely or a.s. when it is clear from the context.

2. Assumptions and estimators

In this section, we describe the Cox regression model with time-varying coefficients, state some assumptions, and define the group SCAD-type and adaptive group Lasso estimator. In deriving the main results, we repeatedly use insights of [17] and [18], of which we also borrow the notation.

Let T and C be a failure time and a censoring time. The interest is in the failure time. However, we observe only $Y = \min\{T, C\}$ on $[0, \tau]$ subject to censoring for some finite τ and $\delta = \mathbf{I}(T \leq C)$. Define

$$N(t) = \delta\, \mathbf{I}(Y \leq t) \quad \text{and} \quad Z(t) = \mathbf{I}(Y \geq t).$$

We also observe a p-dimensional time-dependent covariate $\boldsymbol{X}(t)$. Suppose that $(Y_i, \delta_i, \boldsymbol{X}_i(t))$, where $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^\top$, $i = 1, \ldots, n$, are i.i.d. observations of $(Y, \delta, \boldsymbol{X}(t))$. Then our purpose is to simultaneously carry out variable selection of $\boldsymbol{X}(t)$ and estimation of the time-varying coefficients in (1) below . Assumptions A1-3 on the time-varying Cox regression models reflect a standard setup of Cox regression models. These assumptions and two technical assumptions on $\boldsymbol{X}(t)$ and $Z(t)$ are deferred to section 5.

The hazard function of T_i w.r.t. an appropriate filtration is

$$\lambda(t) = \lambda_0(t) \exp\Big\{ \sum_{j=1}^{p} g_{0j}(t) X_{ij}(t) \Big\} = \lambda_0(t) \exp\{\boldsymbol{g}_0^\top(t) \boldsymbol{X}_i(t)\}, \tag{1}$$

where $\lambda_0(t)$ is an unknown hazard function and $\boldsymbol{g}_0(t) = (g_{01}(t), \ldots, g_{0p}(t))^\top$ is a vector of unknown time-varying coefficients and assumed to be twice continuously differentiable. Details about the sparsity of $\boldsymbol{g}_0(t)$ are given later in this section. Technical assumptions on $\boldsymbol{g}_0(t)$ are postponed to section 5. Note that we do not have to impose any identifiability constraints on the time-varying coefficient $\boldsymbol{g}_0(t)$ since $\boldsymbol{X}(t)$ has no constant element.

We estimate $\boldsymbol{g}_0(t)$ by choosing a basis $\{B_1(t), \ldots, B_{K_n}(t)\}$ on $[0, \tau]$ and maximizing the partial likelihood with/without a penalty term. We allow p to increase moderately (e.g. $p = \mathcal{O}(n^{3/10})$) and consider variable selection.

More precisely for the basis $\{B_1, \ldots, B_{K_n}\}$, we write

$$\boldsymbol{B}(t) = (B_1(t), \ldots, B_{K_n}(t))^\top \quad \text{or} \quad \boldsymbol{B} = (B_1, \ldots, B_{K_n})^\top.$$

Then the covariate vector for the partial likelihood is $\boldsymbol{X}_i(t) \otimes \boldsymbol{B}(t)$. The approximation error ρ_n of the basis $\{B_1, \ldots, B_{K_n}\}$ is defined by

$$\rho_n = \sup_{\boldsymbol{g}_0} \sum_{j=1}^{p} \inf_{\beta_j \in \mathbb{R}^{K_n}} \sup_{0 \leq t \leq \tau} |\beta_j^\top \boldsymbol{B}(t) - g_{0j}(t)|, \tag{2}$$

where $\boldsymbol{g}_0 = (g_{01}, \ldots, g_{0p})^\top$ is over the set of functions satisfying Assumption G in section 5. Then we have $\rho_n = \mathcal{O}(K_n^{-2})$ by the standard theory.

An example of the basis satisfying the following assumption is an equispaced B-spline basis of order $m (m \geq 2)$, see Schumaker [23] for more about B-spline functions.

Assumption B:
(i) $B_j(t)$, $j = 1, \ldots, K_n$, are continuous and bounded on $[0, \tau]$. The upper and lower bounds are allowed to depend on n.
(ii) $\rho_n \to 0$.

Here we define two linear function spaces $\mathbf{G}_0$ and $\mathbf{H}_0$ on $[0, \tau]$ and two norms $\|\cdot\|_\infty$ and $\|\cdot\|_{L_2}$ on them. The ratio of $\|\cdot\|_\infty$ to $\|\cdot\|_{L_2}$ is also considered in (7) below. We define $\mathbf{G}_0$ by

$$\mathbf{G}_0 = \{(\beta_1^\top \boldsymbol{B}(t), \ldots, \beta_p^\top \boldsymbol{B}(t))^\top \,|\, \beta_j \in \mathbb{R}^{K_n},\ j = 1, \ldots, p\}. \tag{3}$$

Let $\mathbf{H}_0$ be the linear space spanned by $\mathbf{G}_0$ and the true coefficient function vector $\boldsymbol{g}_0$. A function space similar to $\mathbf{G}_0$ is defined and called the estimation space in [17]. The dimension of our $\mathbf{G}_0$ is pK_n. Note that the dimension of $\mathbf{G}_0$ in [17] is denoted by N_n and p is fixed there.

For $\boldsymbol{h} = (h_1, \ldots, h_p)^\top \in \mathbf{H}_0$, we define $\|\boldsymbol{h}\|_\infty$ and $\|\boldsymbol{h}\|_{L_2}$ by

$$\|\boldsymbol{h}\|_\infty = \sum_{j=1}^{p} \sup_{0 \leq t \leq \tau} |h_j(t)| \tag{4}$$

and

$$\|\boldsymbol{h}\|_{L_2}^2 = \sum_{j=1}^{p} \|h_j\|_{L_2}^2 = \sum_{j=1}^{p} \frac{1}{\tau} \int_0^\tau h_j^2(t) dt, \tag{5}$$

where we also write for the jth element of $\boldsymbol{h}$,

$$\|h_j\|_{L_2}^2 = \frac{1}{\tau} \int_0^\tau h_j^2(t) dt. \tag{6}$$

The ratio of $\|\cdot\|_\infty$ to $\|\cdot\|_{L_2}$ over $\mathbf{G}_0$ plays an important role and we denote the ratio by A_n.

$$A_n = \sup_{\boldsymbol{g} \in \mathbf{G}_0} \{\|\boldsymbol{g}\|_\infty / \|\boldsymbol{g}\|_{L_2}\}. \tag{7}$$

When we employ an equi-spaced B-spline basis of order $m(m \geq 2)$, we have $A_n \leq C(pK_n)^{1/2}$. The necessary relations between A_n, ρ_n, and pK_n are given in :

Assumption RA:

$$\lim_{n\to\infty} A_n\rho_n = 0 \quad \text{and} \quad \lim_{n\to\infty} n^{-1}A_n^2 \max\{pK_n, \log n\} = 0.$$

When we have $K_n = C_K n^{1/5}$, $\rho_n = \mathcal{O}(K_n^{-2})$, and $A_n = \mathcal{O}\{(pK_n)^{1/2}\}$, Assumption RA implies that $p = o(n^{3/10})$.

In [17], two norms (equivalent to the L_2 norm) are introduced and the norms play an crucial role in the proofs of their main results. We also introduce two similar norms and employ the two norms when we evaluate the eigenvalues of the Hessian matrix of the partial likelihood. Recall that $\mathbf{H}_0$ is spanned by $\mathbf{G}_0$ and $\boldsymbol{g}_0$.

The partial likelihood $l_p(\boldsymbol{h})$ is defined by

$$\begin{aligned} l_p(\boldsymbol{h}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{h}^\top(t)\boldsymbol{X}_i(t)dN_i(t) \\ &\quad - \int_0^\tau \log\Big[n^{-1}\sum_{i=1}^{n} Z_i(t)\exp\Big\{\boldsymbol{h}^\top(t)\boldsymbol{X}_i(t)\Big\}\Big]d\bar{N}(t), \end{aligned} \tag{8}$$

where $\bar{N}(t) = n^{-1}\sum_{i=1}^n N_i(t)$. First we estimate $\boldsymbol{g}_0$ by maximizing $l_p(\boldsymbol{g})$ over $\mathbf{G}_0$ and use the estimator as an initial value of the optimization for the penalized partial likelihood. Define the $\Lambda_p(\boldsymbol{h})$:

$$\begin{aligned} \Lambda_p(\boldsymbol{h}) &= \mathsf{E}\Big\{\int_0^\tau \boldsymbol{h}^\top(t)\boldsymbol{X}(t)dN(t)\Big\} \\ &\quad - \int_0^\tau \log\Big(\mathsf{E}\Big[Z(t)\exp\Big\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t)\Big\}\Big]\Big)d\,\mathsf{E}\{N(t)\}. \end{aligned} \tag{9}$$

For $\boldsymbol{h}_1 \in \mathbf{H}_0$ and $\boldsymbol{h}_2 \in \mathbf{H}_0$, define :

$$(\boldsymbol{h}_1, \boldsymbol{h}_2)_Z(t) = \mathsf{E}[\{\boldsymbol{h}_1^\top(t)\boldsymbol{X}(t)\}\{\boldsymbol{h}_2^\top(t)\boldsymbol{X}(t)\}Z(t)]/\,\mathsf{E}\{Z(t)\} \tag{10}$$

and

$$(\boldsymbol{h}_1, \boldsymbol{h}_2)_{Zn}(t) = \mathsf{E}_n[\{\boldsymbol{h}_1^\top(t)\boldsymbol{X}(t)\}\{\boldsymbol{h}_2^\top(t)\boldsymbol{X}(t)\}Z(t)]/\mathsf{E}_n\{Z(t)\}, \tag{11}$$

where $\mathsf{E}_n(\cdot)$ is the empirical measure, for example, $\mathsf{E}_n\{Z(t)\} = n^{-1}\sum_{i=1}^n Z_i(t)$. As in [17], we define $\|\boldsymbol{h}\|$ for $\boldsymbol{h} \in \mathbf{H}_0$ in terms of the inner product defined in (12) below.

$$(\boldsymbol{h}_1, \boldsymbol{h}_2) = \int_0^\tau (\boldsymbol{h}_1, \boldsymbol{h}_2)_Z(t) d\,\mathsf{E}\{N(t)\}. \tag{12}$$

The empirical version $\|\boldsymbol{h}\|_n$ of $\|\boldsymbol{h}\|$ is defined in terms of the inner product below.

$$(\boldsymbol{h}_1, \boldsymbol{h}_2)_n = \int_0^\tau (\boldsymbol{h}_1, \boldsymbol{h}_2)_{Zn}(t) d\mathsf{E}_n\{N(t)\}. \tag{13}$$

A centered version $\|\boldsymbol{h}\|_0$ of $\|\boldsymbol{h}\|$ is used in [17] for the identifiability constraint. However, we define and use $\|\boldsymbol{h}\|_0$ and the empirical version $\|\boldsymbol{h}\|_{0n}$ only for technical reasons. The centered version is defined in terms of the inner product in (14) below and the empirical version is defined by replacing $\mathsf{E}(\cdot)$ with $\mathsf{E}_n(\cdot)$.

$$\begin{aligned} &(\boldsymbol{h}_1, \boldsymbol{h}_2)_0 \qquad\qquad (14)\\ &= \int_0^\tau [\mathsf{E}\{Z(t)\}]^{-1}\,\mathsf{E}([\boldsymbol{h}_1^\top(t)\boldsymbol{X}(t) - \mathsf{E}\{\boldsymbol{h}_1^\top(t)\boldsymbol{X}(t)Z(t)\}/\,\mathsf{E}\{Z(t)\}] \\ &\quad \times[\boldsymbol{h}_2^\top(t)\boldsymbol{X}(t) - \mathsf{E}\{\boldsymbol{h}_2^\top(t)\boldsymbol{X}(t)Z(t)\}/\,\mathsf{E}\{Z(t)\}]Z(t))d\,\mathsf{E}\{N(t)\}. \end{aligned}$$

The norms defined by (12) and (14) are equivalent to the L_2 norm in (5) and these norms are also equivalent to the empirical counterparts with probability tending to 1. The details are given in Lemmas 1-4 in section 5.

Finally in this section, we define three estimators of $\boldsymbol{g}_0$. The first one is the partial likelihood estimator and defined by

$$\tilde{\boldsymbol{g}}_n = \underset{\boldsymbol{g}\in\mathbf{G}_0}{\operatorname{argmax}}\, l_p(\boldsymbol{g}) \tag{15}$$

It will be shown in Theorem 1 below that the L_2 convergence rate of $\tilde{\boldsymbol{g}}_n$ is :

$$r_{pn} = \max\{(pK_n/n)^{1/2}, \rho_n\}. \tag{16}$$

When we have a moderately large p, it is often the case that only a small number of the covariates are relevant or the model is sparse. Therefore we introduce the sparsity assumption :

Assumption S: For some s, $g_{0j} = 0,\ s+1 \le j \le p$.

To deal with this sparsity, we present two penalized partial likelihoods $Q_p(\boldsymbol{g})$ and $\bar{Q}_p(\boldsymbol{g})$ for $\boldsymbol{g} = (g_1, \ldots, g_p)^\top \in \mathbf{G}_0$.

$$Q_p(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \sum_{j=1}^{p} p_{\lambda_n}(\|g_j\|_{L_2}), \tag{17}$$

where λ_n is a tuning parameter and $p_\lambda(\cdot)$ is a SCAD-type penalty function to be specified in Assumption P below. See (6) for the definition of the L_2 norm. An example of $p_\lambda(\cdot)$ satisfying (i) of Assumption P is the SCAD function. See [13] for the definition of the SCAD function.

Assumption P:
(i) $p_\lambda(t)$ is a monotone increasing and concave function on $[0, \infty)$ with $p_\lambda(0) = 0$. Besides, there are positive constants a_0, b_0, and c_0 such that $p'_\lambda(t) = 0,\ t \geq a_0\lambda$, and $p'_\lambda(t) \geq c_0\lambda,\ 0 < t \leq b_0\lambda$.
(ii) $\lambda_n / r_{pn} \to \infty$ and $\min_{1 \leq j \leq s} \|g_{0j}\|_{L_2} / \lambda_n \to \infty$.

The second assumption in Assumption P means that λ_n should be much larger than the convergence rate and that $\|g_{0j}\|_{L_2}$ should be large enough compared to λ_n.

Another penalized partial likelihood $\bar{Q}_p(\boldsymbol{g})$ is defined by

$$\bar{Q}_p(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \lambda'_n \sum_{j=1}^{p} w_j \|g_j\|_{L_2}, \tag{18}$$

where λ'_n is another tuning parameter and w_j, $j = 1, \ldots, p$, are weights to be constructed from a preliminary estimator. Notice that $\bar{Q}_p(\boldsymbol{g})$ is a concave function.

Finally the group SCAD-type estimator $\hat{\boldsymbol{g}}_n$ and the adaptive group Lasso estimator $\bar{\boldsymbol{g}}_n$ are given by

$$\hat{\boldsymbol{g}}_n = \operatorname*{argmax}_{\boldsymbol{g} \in \mathbf{G}_0} Q_p(\boldsymbol{g}) \quad \text{and} \quad \bar{\boldsymbol{g}}_n = \operatorname*{argmax}_{\boldsymbol{g} \in \mathbf{G}_0} \bar{Q}_p(\boldsymbol{g}). \tag{19}$$

We can also define $l_s(\boldsymbol{g})$, $Q_s(\boldsymbol{g})$, and $\bar{Q}_s(\boldsymbol{g})$ for the s in Assumption S by ignoring the last $(p - s)$ elements of the covariates or taking $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{is}(t))^\top$ and $\boldsymbol{g} = (g_1, \ldots, g_s)^\top$.

The main contribution here is to prove that we can select the relevant covariates consistently by using $\hat{\boldsymbol{g}}_n$ or $\bar{\boldsymbol{g}}_n$. Besides, both of them achieve the rate of convergence r_{sn}, where $r_{sn} = \max\{(sK_n/n)^{1/2}, \rho_n\}$ and r_{sn} is the convergence rate we obtain by maximizing $l_s(\boldsymbol{g})$. Besides, we establish the oracle property of the SCAD-type estimator. We discuss how to compute $\tilde{\boldsymbol{g}}_n$, $\hat{\boldsymbol{g}}_n$, and $\bar{\boldsymbol{g}}_n$ in section 4.

3. Main theorems

The L_2 convergence rate of the partial likelihood estimator $\tilde{\boldsymbol{g}}_n$ is derived in Theorem 1. The properties of the group SCAD-type and the adaptive group Lasso estimator are considered in Theorems 2-3 and Theorems 4-5, respectively. We also comment on the semi-varying coefficient model in Remark 2. Recall that Assumptions A1-3, X, M, and G are given later in section 5.

Theorem 1. *Suppose that Assumptions A1-3, X, M, G, B, and RA hold and $r_{pn} \to 0$. Then with probability tending to 1, there is a unique maximizer $\tilde{\boldsymbol{g}}_n = (\tilde{g}_{n1}, \ldots, \tilde{g}_{np})^\top$ of $l_p(\boldsymbol{g})$ over $\mathbf{G}_0$ and we have*

$$\|\tilde{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} = \mathcal{O}_p(r_{pn}).$$

The existence of the group SCAD-type estimator is verified in Theorem 2 and the sparsity and oracle property is established is Theorem 3.

Theorem 2. *Suppose that all the assumptions in Theorem 1 and Assumptions P and S hold. Then for any positive ϵ, there is a positive constant M such that*

$$\lim_{n\to\infty} \mathsf{P}(\ \textit{There is a local maximizer } \hat{\boldsymbol{g}}_n \\ \textit{of } Q_p(\boldsymbol{g}) \textit{ over } \mathbf{G}_0 \textit{ such that } \|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} \le M r_{pn}) > 1 - \epsilon.$$

Before we present Theorem 3, we define two properties. If the local maximizer of $\hat{\boldsymbol{g}}_n = (\hat{g}_{n1}, \ldots, \hat{g}_{np})^\top$ satisfies under Assumption S,

$$\hat{g}_{nj} = 0,\ j = s+1, \ldots, p, \tag{20}$$

with probability tending to 1, we say that $\hat{\boldsymbol{g}}_n$ has the sparsity. The maximizer of $l_s(\boldsymbol{g})$ is called an oracle estimator since we use the knowledge of the true model under Assumption S. If an estimator is asymptotically equivalent to such an oracle estimator, we say that the estimator has the oracle property.

It is known that (ii) easily follows from (i) in Theorem 3 below due to the flatness of $p_\lambda(t)$ on $[a_0\lambda, \infty)$. For example, see Fan and Lv [14].

Theorem 3. *Suppose that the assumptions in Theorem 2 hold and let $\{d_n\}$ be a sequence of positive numbers satisfying $d_n \to \infty$, $\lambda_n/(d_n r_{pn}) \to \infty$, and $A_n d_n r_{pn} = \mathcal{O}(1)$.*

(i) *With probability tending to 1, any local maximizer* $\hat{\boldsymbol{g}}_n$ *of* $Q_p(\boldsymbol{g})$ *over* $\mathbf{G}_0$ *such that* $\|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} \leq d_n r_{pn}$ *satisfies (20).*
(ii) *With probability tending to 1, the local maximizer in* (i) *is the unique maximizer of* $l_s(\boldsymbol{g})$ *and satisfies*

$$\sum_{j=1}^{s} \|\hat{g}_{nj} - g_{0j}\|_{L_2}^2 = \mathcal{O}(r_{sn}^2).$$

In Theorems 4 and 5 below, we state the properties of the adaptive group Lasso estimator. We comment on how to choose the weights $\{w_j\}$ in Remark 1 below.

Theorem 4. *Suppose that the assumptions in Theorem 1 and Assumption S hold and that* $\lambda_n' \sqrt{s} \max_{1 \leq j \leq s} w_j / r_{pn} = \mathcal{O}_p(1)$. *Then with probability tending to 1, there is a unique maximizer* $\bar{\boldsymbol{g}}_n = (\bar{g}_{n1}, \ldots, \bar{g}_{np})^\top$ *of* $\bar{Q}_p(\boldsymbol{g})$ *over* $\mathbf{G}_0$ *and we have*

$$\|\bar{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} = \mathcal{O}_p(r_{pn}).$$

Remark 1. Suppose we take $w_j = 1/\|\tilde{g}_{nj}\|_{L_2}$. The assumption on $\{w_j\}$ in Theorem 4 is satisfied if we have $\lambda_n' \sqrt{s} / (r_{pn} \min_{1 \leq j \leq s} \|g_{0j}\|_{L_2}) = \mathcal{O}(1)$ and $\min_{1 \leq j \leq s} \|g_{0j}\|_{L_2} / r_{pn} \to \infty$, If we also have $\lambda_n' / r_{pn}^2 \to \infty$, the assumptions on $\{w_j\}$ in Theorem 5 below are also satisfied.

Theorem 5. *Suppose that the assumptions in Theorem 4 hold and that* $\lambda_n' \min_{s < j \leq p} w_j / (r_{pn}) \to \infty$ *in probability and* $A_n r_{pn} = \mathcal{O}(1)$. *Then with probability tending to 1, the unique maximizer* $\bar{\boldsymbol{g}}_n$ *has the sparsity and is equal to the unique maximizer of* $\bar{Q}_s(\boldsymbol{g})$. *In addition we have*

$$\sum_{j=1}^{s} \|\bar{g}_{nj} - g_{0j}\|_{L_2}^2 = \mathcal{O}_p(r_{sn}^2).$$

We have demonstrated that the group SCAD estimator has the sparsity and oracle property. As for the adaptive group Lasso estimator, we have established the sparsity and the improved L_2 convergence rate. Generally speaking, it is difficult to prove the oracle property of the adaptive group Lasso estimator due to the property of the Lasso penalty function. On the other hand, the uniqueness easily follows from the concavity of $\bar{Q}_p(\boldsymbol{g})$. When some statistical inference is necessary, we recommend to ignore the influences

of the penalty term and use the inverse of the observed Fisher information matrix of the partial likelihood as the estimate of the variance. The validity of this procedure is a topic of future research.

The condition on p is rather restrictive due to Assumption RA. When we deal with the case of a larger p we will have to use the methodology of [2] by imposing much more restrictive assumptions on the model and the properties of covariates. One of the main purposes of this paper is to establish the desirable properties of the estimators under simple, mild, and interpretable assumptions. See assumptions in section 5.

Finally we comment on how to select a semi-varying coefficient model from the varying coefficient model.

Remark 2. Suppose that the true model is a semi-varying coefficient model. Then we can detect the semi-varying coefficient model with probability tending one by modifying the estimators in the following way. We decompose g_j of $\boldsymbol{g} = (g_1, \ldots, g_p) \in \mathbf{G}_0$, by

$$g_j(t) = \frac{1}{\tau}\int_0^\tau g_j(s)ds + \Big\{g_j(t) - \frac{1}{\tau}\int_0^\tau g_j(s)ds\Big\} = g_{aj} + g_{bj}(t)$$

and

$$\|g_j\|_{L_2}^2 = |g_{aj}|^2 + \|g_{bj}\|_{L_2}^2.$$

Then we define $Q_p'(\boldsymbol{g})$ and $\bar{Q}_p'(\boldsymbol{g})$ by

$$Q_p'(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \sum_{j=1}^p \{p_{\lambda_n}(|g_{aj}|) + p_{\lambda_n}(\|g_{bj}\|_{L_2})\}$$

and

$$\bar{Q}_p'(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \lambda_n' \sum_{j=1}^p (w_{1j}|g_{aj}| + w_{2j}\|g_{bj}\|_{L_2}),$$

where w_{1j} and w_{2j} are weights. Note that $\bar{Q}_p'(\boldsymbol{g})$ is similar to (6) of [26]. See also Zhang et al. [31]. We derive almost the same results as in Theorems 2-5 for $Q_p'(\boldsymbol{g})$ and $\bar{Q}_p'(\boldsymbol{g})$ in the same way because the decomposition of g_j into g_{aj} and g_{bj} does not depend on data and the decomposition is the orthogonal

one. This implies that we select only the g_{aj} component consistently for any j such that $\|g_j\|_{L_2} \neq 0$ and g_j has no nonlinear component g_{bj}. Therefore we can detect the true semi-varying coefficient model from the varying coefficient model consistently. However, the convergence rate is still r_{sn} even for the linear component. We will have to use a two-step estimator to improve the convergence rate of the linear component or employ another proof to show that the proposed estimators have the improved convergence rate for the linear component. This remark also applies to the model in section 6.

4. Simulation studies and a real example

We carried out some simulations by using R to examine the finite sample properties of the group SCAD-type and adaptive group Lasso estimators in (19). We considered two models whose hazard functions are given by

$$h_1(t) = \lambda_0 \exp\{-X_1 + X_2 w_1 \log(t+1)\}$$

and

$$h_2(t) = \lambda_0 \exp(-X_1 + X_2 w_2 t),$$

where $\lambda_0 = 0.275$,

$$\frac{w_1^2}{\tau}\int_0^\tau \{\log(t+1)\}^2 dt = 1, \quad \text{and} \quad \frac{w_2^2}{\tau}\int_0^\tau t^2 dt = 1.$$

We call them Model 1 and Model 2, respectively. We follow [1] in generating survival times and took $\tau = 4$, $p = 14$, $K_n = 5$, and $n = 600$ in the simulations. Only X_1 and X_2 are relevant. The replication number is 100 because only one iteration takes several minutes and the standard errors of the simulated results are small enough. The covariates $X_1, \ldots, X_{14}$ follow $U(0,2)$ independently of each other and the censoring time C follows $\frac{1}{2}U(0,\tau) + \frac{1}{2}\mathbf{I}(C = \tau)$ independently of the covariates. The censoring rate is about 50% in both models.

We employ the coxph function to compute the partial likelihood estimator $\tilde{\boldsymbol{g}}_n$ in (15). As for the basis function, we chose an equi-spaced quadratic B-spline basis $\boldsymbol{B} = (B_1, \ldots, B_{K_n})^\top$ with $K_n = 5$. We represent g_j in $\boldsymbol{g} = (g_1, \ldots, g_p)^\top$ as $g_j = \beta_j^\top \boldsymbol{B}$ and define an pK_n vector $\boldsymbol{\beta}$ by $\boldsymbol{\beta} = (\beta_1^\top, \ldots, \beta_{K_n}^\top)^\top$ to describe the algorithms.

We approximate $l_p(\boldsymbol{g})$ in a neighborhood of $\boldsymbol{g}^0 = (g_1^0, \ldots, g_p^0)^\top$ by

$$l_p(\boldsymbol{g}^0) + (\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top \frac{\partial l_p}{\partial \boldsymbol{\beta}}(\boldsymbol{g}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^\top \frac{\partial^2 l_p}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}(\boldsymbol{g}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0), \tag{21}$$

where $g_j^0 = \beta_j^{0\top} \boldsymbol{B}$, $j = 1, \ldots, p$, and $\boldsymbol{\beta}^0 = (\beta_1^{0\top}, \ldots, \beta_p^{0\top})^\top$.

We give the details of the computation of the group SCAD-type estimator. We use the SCAD function with $a = 3.7$ and approximate $p_{\lambda_n}(\|g_j\|_{L_2})$ in a neighborhood of g_j^0 by

$$p_{\lambda_n}(\|g_j^0\|_{L_2}) + \frac{p'_{\lambda_n}(\|g_j^0\|_{L_2})}{2\|g_j^0\|_{L_2}}(\|g_j\|_{L_2}^2 - \|g_j^0\|_{L_2}^2) \tag{22}$$

as in [13]. The computational algorithm is as follows:

1. Put $\hat{\boldsymbol{g}}^{(1)} = \tilde{\boldsymbol{g}}_n$, where $\tilde{\boldsymbol{g}}_n$ is defined in (15).
2. Approximate $Q_p(\boldsymbol{g})$ in a neighborhood of $\hat{\boldsymbol{g}}^{(m)}$ by using (21) and (22) and minimize the approximation.
3. Denote the solution in step 2 by $\hat{\boldsymbol{g}}^{(m+1)} = (\hat{g}_1^{(m+1)}, \ldots, \hat{g}_p^{(m+1)})^\top$. Replace $\hat{g}_j^{(m+1)}$ with 0 and remove the j from step 2 in the iteration once $\|\hat{g}_j^{(m+1)}\|_{L_2} \leq 0.01$.
4. Iterate steps 2 and 3 until $\|\hat{\boldsymbol{g}}^{(m)} - \hat{\boldsymbol{g}}^{(m+1)}\|_{L_2} \leq 0.005$. Let q be the number of the finally selected covariates.
5. Compute the partial likelihood estimator with the finally selected covariates and denote the partial likelihood estimator by $\tilde{\boldsymbol{g}}_{qn} = (\tilde{g}_{qn1}, \ldots, \tilde{g}_{qnp})^\top$. We put $\tilde{g}_{qnj} = 0$ if j is not selected in step 4.

We take $\lambda_n = b(pK_n/n)^{1/2}$ with $b = 0.2, 0.3, 0.4$ and select the one with the smallest BIC as the group SCAD estimator, where

$$BIC = -2 \log l_p(\tilde{\boldsymbol{g}}_{qn}) + qK_n \log n.$$

We can define AIC by replacing $qK_n \log n$ with $2qK_n$. When the number of the covariates is large, it is very time-consuming and impractical to compute the BIC or AIC for all the submodels. It will be very useful to the procedures here and the information criteria simultaneously. We adopt the above definitions of the information criteria and do not enter into the controversy although there is a controversy on how to define the information criteria in the case of the partial likelihood.

We have $\tilde{\boldsymbol{g}}_{qn} = \hat{\boldsymbol{g}}_n$ in most replications of the simulations and the average iteration number is 5.8 for selected λ_n.

Next we describe the computation of the adaptive group Lasso estimator. In fact we replaced $\|g_j\|_{L_2}$ with $|\beta_j|/\sqrt{K_n}$ in (18) as in [26], where $g_j = \beta_j^\top \boldsymbol{B}$, for computational simplicity. Note that Theorems 4 and 5 about the adaptive group Lasso estimator still hold because $\|g_j\|_{L_2}$ is equivalent to $|\beta_j|/\sqrt{K_n}$. We also followed the algorithm in [26] except that we used $\mathbb{X}$ in the iteration such that $\mathbb{X}^2 = H$ and $\mathbb{X}$ is symmetric and that we chose the weights as in Remark 1. They used the Cholesky decomposition of H for $\mathbb{X}$ in [26]. They used the second order approximation in (21) and the KKT condition when they proposed their algorithm. See [26] for the details of the algorithm. We apply the same convergence criterion as for the group SCAD estimator. We take $\lambda_n' = b'pK_n/n$ with $b' = 0.1, 0.3, 0.5$ and select the one with the smallest GCV. In these simulations, we adopt the same GCV as in [26]. The average iteration number is 6.5 for selected λ_n'.

We present the simulation results in Tables 1-4. We define the mean integrated squared error, MISE, of $\hat{g}_j$ by

$$\mathsf{E}\Big[\frac{1}{\tau}\int_0^\tau \{\hat{g}_j(t) - g_{0j}(t)\}^2 dt\Big].$$

The MISE of the other estimators is similarly defined. We define IMISE, MISE, and PMISE in Tables 1 and 2 as follows:

IMISE : The MISE of the initial partial likelihood estimator $\tilde{g}_j$
MISE : The MISE of the group SCAD estimator $\hat{g}_j$ or adaptive group Lasso estimator $\bar{g}_j$
PMISE : The MISE of the partial likelihood estimator $\tilde{g}_{qnj}$, which is computed with the finally selected covariates

The numbers in parentheses are standard errors in Tables 1 and 2. Note that MISE's of the adaptive group Lasso estimators are large. They are much smaller when we choose and fix a smaller λ_n', for example, $b = 0.15$.

Table 1: MISE's of Model 1

		IMISE	MISE	PMISE
$j=1$	SCAD	0.111(0.004)	0.082(0.003)	0.082(0.003)
	Lasso	0.111(0.004)	0.187(0.003)	0.079(0.003)
$j=2$	SCAD	0.108(0.004)	0.075(0.002)	0.075(0.002)
	Lasso	0.108(0.004)	0.267(0.004)	0.074(0.002)

Table 2: MISE's of Model 2

		IMISE	MISE	PMISE
$j=1$	SCAD	0.095(0.003)	0.075(0.002)	0.075(0.002)
	Lasso	0.095(0.003)	0.158(0.003)	0.074(0.002)
$j=2$	SCAD	0.103(0.004)	0.068(0.002)	0.068(0.002)
	Lasso	0.103(0.004)	0.232(0.004)	0.067(0.002)

Both estimators selected only the relevant covariates X_1 and X_2. In every replication of Models 1 and 2, X_1 and X_2 were selected by the the group SCAD and adaptive group Lasso estimators. Only fifteen and no irrelevant covariates were falsely selected among the total 100 replications of Model 1 by the the group SCAD and adaptive group Lasso estimator, respectively. Only six and two irrelevant covariates were falsely selected respectively in the case of Model 2. In Tables 3 and 4, we present the mean of the number of falsely selected covariates to show how the tuning parameters affect these estimators.

Table 3: Numbers of falsely selected covariates (SCAD)

b	0.2	0.3	0.4
Model 1	6.00	1.10	0.15
Model 2	5.34	0.64	0.06

Table 4: Numbers of falsely selected covariates (Lasso)

b	0.2	0.3	0.4
Model 1	1.98	0.03	0.00
Model 2	1.54	0.02	0.00

We have the following implications from Tables 1-4.

1. Tables 3 and 4 show that selection of tuning parameters λ_n and λ_n' is critical to variable selection.
2. The group SCAD estimator is equal to the partial likelihood estimator with the selected variables in most of the replications.

3. In Tables 1 and 2, the MISE's of the adaptive group Lasso estimator are much larger than those of the initial partial likelihood estimator. This means that selection of the tuning parameter by GCV may not work well or we should use the adaptive group Lasso estimator only for variable selection. The GCV criterion tends to choose a larger λ_n' and cause a larger bias. On the other hand, the tuning parameter selection by BIC works well for the group SCAD estimator.

The simulation studies suggest that the group SCAD estimator with BIC tuning parameter selection works well. We know we should conduct more extensive simulation studies to obtain a conclusion about the tuning parameter selection. However, it takes a lot of time to compute these estimators and the simulation studies of this paper are limited because of the computational time.

Next we present a real example. As in [26], we apply the above two procedures to the well-known PBC (primary biliary cirrhosis) data. The data is often used in the literature of time-varying Cox regression models. PBC is a fatal liver disease and the data is from a trial of comparing the drug D-penicillamine and a placebo. Times to death or censoring are recorded. The details are given in Fleming and Harrington [15] and the survival package of R. We use the first 312 randomized cases and consider only ten covariates among 17 covariates in the data set for numerical stability of the coxph function in the survival package. We remove two cases with missing covariates and our sample size is 310. In this study, we consider the following covariates. We normalize continuous covariates so that the mean is 0 and the variance is 1.

1) treatment indicator (0:placebo, 1:D-penicillamine); 2) normalized age; 3) sex (0:male, 1:female); 4) presence of hepatomegaly (0:no, 1:yes); 5) presence of edema (0, 0.5, 1 according to the severity); 6) normalized log serum bilirubin; 7) normalized serum albumin; 8) normalized urine copper; 9) normalized log prothrombin time; 10) histologic stage of disease (0, 1/3, 2/3, 1)

In this study, we have $n = 310$, $p = 10$, and $\tau = 12$ years. We take $K_n = 4$, $\lambda_n = b(pK_n/n)^{1/2}$ with $b = 0.4, 0.5, 0.6, 0.7$ for the group SCAD estimator, and $\lambda_n' = b'pK_n/n$ with $b' = 0.2, 0.3, 0.4, 0.5, 0.6$ for the adaptive group Lasso estimator. We finally select the variables by using AIC and BIC for the group SCAD estimator and AIC, BIC, and GCV for the adaptive group Lasso estimator.

We present the results of variable selection in Table 5. In the table, we also

give the squared L_2 norms of estimated functions $\|\tilde{g}_{nj}\|^2_{L_2}$ for the initial PL estimators. For the SCAD and Lasso estimators, we reestimate the coefficient functions for selected variables by employing the coxph function and present the squared L_2 norms. Note that we use the two procedures only for variable selection.

Table 5: Selected variables and the squared L_2 norms

	Initial PLE	SCAD BIC	SCAD AIC	Lasso BIC	Lasso GCV	Lasso AIC
1) treatment	0.596	0	0.569	0	0	0
2) age	0.502	0	0.509	0.263	0.263	0.259
3) sex	0.616	0	0.509	0	0	0
4) hepato	0.690	0	1.147	0	0	0
5) edema	3.320	1.636	2.876	1.708	1.708	1.542
6) bilirubin	4.676	4.153	3.638	4.134	4.134	4.749
7) albumin	0.841	1.202	0.806	0.650	0.650	0.655
8) copper	0.117	0	0	0	0	0.090
9) prothrombin	0.200	0.314	0.234	0.248	0.248	0.265
10) stage	0.694	0	1.237	0.610	0.610	0.621

Comparing the results here to those in [26], we notice that the urine copper is selected only in the case of Lasso and AIC and that only the Lasso and AIC result coincides with those in [26]. BIC tends to choose smaller sets of covariates due to its penalty term.

We need more extensive simulation studies and real data analysis to examine the finite sample properties of the procedures. It is a topic of future research.

5. Proofs of Theorems 1-5

We prove Theorems 1-5 in this section. First we describe some technical assumptions. Next we state Lemmas 1-4 and Propositions 1-4. Finally we prove the theorems by using the propositions. The proofs of the lemmas and propositions are postponed to section 7. We prove the theorems by following the methodology in [17] and Propositions 1, 2, and 3 correspond to their Lemma 7, Lemma 10, and Lemma 11, respectively.

The assumptions A1-3 below are about the Cox regression model with time-varying coefficients.

Assumption A1: There is a suitable filtration $\{\mathcal{F}_t\}$ such that $Y_i(t)$ are adapted to $\{\mathcal{F}_t\}$ and $(Z_i(t), \boldsymbol{X}_i(t))$ is predictable w.r.t. $\{\mathcal{F}_t\}$.

Assumption A2: When we have no censoring time, the hazard function of T_i w.r.t. $\{\mathcal{F}_t\}$ is given by

$$\lambda_i(t) = \lambda_0(t) \exp\Big\{ \sum_{j=1}^{p} g_{0j}(t) X_{ij}(t) \Big\} = \lambda_0(t) \exp\{\boldsymbol{g}_0^\top(t) \boldsymbol{X}_i(t)\},$$

where $\lambda_0(t)$ is an unknown hazard function and $\boldsymbol{g}_0(t) = (g_{01}(t), \ldots, g_{0p}(t))^\top$ is a vector of unknown time-varying coefficients.

Assumption A3: The censoring time C_i satisfies the independent censoring condition. This means that the compensator of $N_i(t)$ w.r.t. $\{\mathcal{F}_t\}$ is equal to

$$\int_0^t Z_i(s)\lambda_i(s)ds = \int_0^t Z_i(s) \exp\{\boldsymbol{g}_0^\top(s) \boldsymbol{X}_i(s)\} \lambda_0(s) ds. \tag{23}$$

Then

$$M_i(t) = N_i(t) - \int_0^t Z_i(s)\lambda_i(s)ds \tag{24}$$

is a martingale process w.r.t. $\{\mathcal{F}_t\}$.

We need following technical assumptions on $\boldsymbol{X}(t)$ and $Z(t)$. We set

$$\Sigma(t) = \mathrm{Var}(\boldsymbol{X}(t)) \quad \text{and} \quad \Omega(t) = \mathsf{E}\{\boldsymbol{X}(t)\boldsymbol{X}^\top(t)\}.$$

Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalue of a symmetric matrix A.

Assumption X:
(i) There are positive constants C_m and C_M such that uniformly in t in $[0, \tau]$, $C_m \le \lambda_{\min}(\Sigma(t)) \le \lambda_{\min}(\Omega(t))$ and $\lambda_{\max}(\Sigma(t)) \le \lambda_{\max}(\Omega(t)) \le C_M$.
(ii) $X_{ij}(t)$ is uniformly bounded in i, j, and t a.s.

The first assumption in Assumption X is easy to check and does not depend on the sample property of $\boldsymbol{X}(t)$. Besides, the density function of $\boldsymbol{X}(t)$ is not necessary. The second one is necessary because we need to evaluate $\exp\{\boldsymbol{g}_0^\top(t) \boldsymbol{X}_i(t)\}$.

Assumption M:

(i) There is a positive constant C_Z such that $\mathsf{E}\{Z(t)|\boldsymbol{X}(t)\} \geq C_Z$ on $[0,\tau]$ a.s.
(ii) There are positive constants C_{L1} and C_{L2} such that $C_{L1} \leq \lambda_0(t) \leq C_{L2}$ on $[0,\tau]$.

The first assumption in Assumption M is a standard assumption and the second one is necessary since we deal with time-varying coefficient models on $[0,\tau]$.

We describe the assumptions on $\boldsymbol{g}_0(t) = (g_{01}(t),\ldots,g_{0p}(t))^\top$ on $[0,\tau]$.

Assumption G:
(i) $g_{0j}(t)$, $j = 1,\ldots,p$, are twice continuously differentiable on $[0,\tau]$ and may depend on the sample size n.
(ii) There are positive constants C_{g0} and C_{g2} such that

$$\sum_{j=1}^{p} \sup_{0\leq t\leq\tau} |g_{0j}(t)| \leq C_{g0} \quad \text{and} \quad \sum_{j=1}^{p} \sup_{0\leq t\leq\tau} |g''_{0j}(t)| \leq C_{g2}.$$

We impose the first assumption in Assumption G for simplicity of presentation and the second one is necessary when we evaluate $\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}_i(t)\}$ and define the approximation error ρ_n.

We state the important properties of the norms introduced in section 2. The proofs are postponed to section 7. Similar results are given in [17]. We write $a_n \sim b_n$ when $a_n \leq C_1 b_n$ and $a_n \geq C_2 b_n$ for some positive constants C_1 and C_2. The first two lemmas are about the equivalences between the norms and the L_2 norm. The last two lemmas evaluates the differences between the norms and the empirical counterparts.

Lemma 1. *Suppose that Assumptions A1-3, X, M, B, and G hold. Then uniformly in* $\boldsymbol{h} \in \mathbf{H}_0$,

$$\|\boldsymbol{h}\| \sim \|\boldsymbol{h}\|_{L_2}.$$

Lemma 2. *Suppose that Assumptions A1-3, X, M, B, and G hold. Then uniformly in* $\boldsymbol{h} \in \mathbf{H}_0$,

$$\|\boldsymbol{h}\|_0 \sim \|\boldsymbol{h}\|_{L_2}.$$

Lemma 3. *Suppose that Assumptions A1-3, X, M, B, G, and RA hold. Then*

$$\sup_{\boldsymbol{g}_1,\boldsymbol{g}_2\in\mathbf{G}_0} \left| \frac{(\boldsymbol{g}_1,\boldsymbol{g}_2)_n - (\boldsymbol{g}_1,\boldsymbol{g}_2)}{\|\boldsymbol{g}_1\|\|\boldsymbol{g}_2\|} \right| = o_p(1).$$

Lemma 4. *Suppose that Assumptions A1-3, X, M, B, G, and RA hold. Then*

$$\sup_{\boldsymbol{g}_1,\boldsymbol{g}_2\in\mathbf{G}_0}\Big|\frac{(\boldsymbol{g}_1,\boldsymbol{g}_2)_{0n}-(\boldsymbol{g}_1,\boldsymbol{g}_2)_0}{\|\boldsymbol{g}_1\|_0\|\boldsymbol{g}_2\|_0}\Big|=o_p(1).$$

Proposition 1. *Suppose that the same assumptions hold as in Theorem 1. Then there is a unique maximizer* $\boldsymbol{g}_n^*=(g_{n1}^*,\ldots,g_{np}^*)^\top$ *of* $\Lambda_p(\boldsymbol{g})$ *over* $\mathbf{G}_0$ *satisfying* $\|\boldsymbol{g}_n^*-\boldsymbol{g}_0\|=\mathcal{O}(\rho_n)$.

By Assumptions G and RA and Lemma 1 we have $\|\boldsymbol{g}_n^*\|_\infty\le C_\infty$ for some C_∞.

Here we choose an orthonormal basis $\{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{pK_n}\}$ of $\mathbf{G}_0$ w.r.t. $\|\cdot\|$ and define an $R^{p\times(pK_n)}$-valued function $\Phi(t)$ on $[0,\tau]$ by

$$\Phi(t)=(\boldsymbol{\phi}_1(t),\ldots,\boldsymbol{\phi}_{pK_n}(t))\quad\text{or}\quad\Phi=(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{pK_n}).\tag{25}$$

This orthonormal basis is just a technical tool as in [17] and the results do not depend on any particular choice.

By using this basis, we can represent $\boldsymbol{g}_n^*$ for some $\boldsymbol{\beta}_n^*\in R^{pK_n}$ as

$$\boldsymbol{g}_n^*=\Phi\boldsymbol{\beta}_n^*.\tag{26}$$

When $\boldsymbol{g}=\Phi\boldsymbol{\beta}$, $l_p(\boldsymbol{g})=l_p(\Phi\boldsymbol{\beta})$ is represented as

$$\begin{aligned}l_p(\Phi\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n\int_0^\tau\{\Phi(t)\boldsymbol{\beta}\}^\top\boldsymbol{X}_i(t)dN_i(t)\\ &\quad-\int_0^\tau\log\Big(n^{-1}\sum_{i=1}^n Z_i(t)\exp\Big[\{\Phi(t)\boldsymbol{\beta}\}^\top\boldsymbol{X}_i(t)\Big]\Big)d\bar{N}(t).\end{aligned}$$

and we write

$$S_p(\boldsymbol{\beta})=\frac{\partial l_p}{\partial\boldsymbol{\beta}}(\Phi\boldsymbol{\beta})\quad\text{and}\quad D_p(\boldsymbol{\beta})=\frac{\partial^2 l_p}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}(\Phi\boldsymbol{\beta}).\tag{27}$$

We evaluate $S_p(\boldsymbol{\beta}_n^*)$ in Proposition 2 below and evaluate the eigenvalues of $D_p(\boldsymbol{\beta})$ in Proposition 3 below. Proposition 3 is stated in a more general form than in [17].

Proposition 2. *Suppose that the same assumptions hold as in Theorem 1. Then we have*

$$|S_p(\boldsymbol{\beta}_n^*)|=\mathcal{O}_p\Big\{\Big(\frac{pK_n}{n}\Big)^{1/2}\Big\}.$$

Proposition 3. *Suppose that the same assumptions hold as in Theorem 1. When $\|\Phi\boldsymbol{\beta}\|_\infty \leq M$ for some positive M, there are positive constants M_1 and M_2 such that*

$$-M_1 \leq \lambda_{\min}(D_p(\boldsymbol{\beta})) \leq \lambda_{\max}(D_p(\boldsymbol{\beta})) \leq -M_2$$

uniformly in $\boldsymbol{\beta}$ with probability tending to 1.

Proposition 4 below is necessary to the proof of the sparsity.

Proposition 4. *Suppose that the same assumptions hold as in Theorem 1. When $\|\Phi\boldsymbol{\beta}_1\|_\infty \leq M$ for some positive M, we have*

$$(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top S_p(\boldsymbol{\beta}_1) = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top S_p(\boldsymbol{\beta}_n^*) + \mathcal{O}_p(|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_n^*|)$$

uniformly in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ with probability tending to 1.

Now we start to prove Theorems 1-5.

Proof of Theorem 1. We have only to show that there is a unique maximizer $\tilde{\boldsymbol{\beta}}_n$ of $l_p(\Phi\boldsymbol{\beta})$ over $\mathbf{G}_0$ such that $\|\Phi\tilde{\boldsymbol{\beta}}_n - \Phi\boldsymbol{\beta}_n^*\| = \mathcal{O}_p(r_{pn})$ with probability tending to 1. Then the desired result follows from Lemma 1 and Proposition 1.

Define Γ_M for a positive constant M by

$$\Gamma_M = \{\boldsymbol{g} = \Phi\beta \,|\, |\boldsymbol{\beta} - \boldsymbol{\beta}_n^*| = M(pK_n/n)^{1/2}\}$$

and consider the Taylor expansion of $l_p(\Phi\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_n^*$.

$$\begin{aligned} l_p(\Phi\boldsymbol{\beta}) &= l_p(\Phi\boldsymbol{\beta}_n^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}_n^*)^\top S_p(\boldsymbol{\beta}_n^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}_n^*)^\top D_p(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}_n^*) \quad (28) \\ &= l_p(\Phi\boldsymbol{\beta}_n^*) + J_1(\boldsymbol{\beta}) + J_2(\boldsymbol{\beta}), \end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_n^*$. By Proposition 2, we have uniformly on Γ_M,

$$J_1(\boldsymbol{\beta}) = \frac{M(pK_n)}{n}\mathcal{O}_p(1). \quad (29)$$

By Proposition 1 and Assumption RA, we have

$$\|\Phi\boldsymbol{\beta}\|_\infty \leq \|\Phi\boldsymbol{\beta}_n^*\|_\infty + A_n CM(pK_n/n)^{1/2} \leq C. \quad (30)$$

We have by Proposition 3 and (30) that for some positive constant M_2,

$$J_2(\boldsymbol{\beta}) \leq -M^2 M_2 p K_n / n \tag{31}$$

uniformly on Γ_M with probability tending to 1. Combining (28), (29), and (31), we obtain

$$\lim_{M\to\infty} \lim_{n\to\infty} \mathsf{P}\Big(\sup_{\Phi\boldsymbol{\beta}\in\Gamma_M} l_p(\Phi\boldsymbol{\beta}) < l_p(\Phi\boldsymbol{\beta}_n^*)\Big) = 1. \tag{32}$$

The concavity of $l_p(\Phi\boldsymbol{\beta})$, Proposition 3, and (32) imply that there is a unique maximizer $\tilde{\boldsymbol{\beta}}_n$ of $l_p(\Phi\boldsymbol{\beta})$ over $\mathbf{G}_0$ such that $\|\Phi\tilde{\boldsymbol{\beta}}_n - \Phi\boldsymbol{\beta}_n^*\| = \mathcal{O}_p(r_{pn})$. Hence the proof of Theorem 1 is complete.

Before the proof of Theorem 2, we define $\bar{\boldsymbol{g}}_n^*$ by

$$\bar{\boldsymbol{g}}_n^* = (g_{n1}^*, \ldots, g_{ns}^*, 0, \ldots, 0)^\top. \tag{33}$$

Recall that $\boldsymbol{g}_n^*$ is given in Proposition 1. Proposition 1 implies that $\|\bar{\boldsymbol{g}}_n^* - \boldsymbol{g}_0\| = \mathcal{O}(\rho_n)$ under Assumption S.

Proof of Theorem 2. We have only to demonstrate that for any positive ϵ, there is a positive constant M such that

$$\lim_{n\to\infty} \mathsf{P}\Big(\sup_{\boldsymbol{g}\in\bar{\Gamma}_M} Q_p(\boldsymbol{g}) < Q_p(\bar{\boldsymbol{g}}_n^*)\Big) > 1 - \epsilon, \tag{34}$$

where $\bar{\Gamma}_M = \{\boldsymbol{g} = \Phi\boldsymbol{\beta} \,|\, |\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_n^*| = M r_{pn}\}$ and $\bar{\boldsymbol{g}}_n^* = \Phi\bar{\boldsymbol{\beta}}_n^*$.

Write

$$\begin{aligned} Q_p(\boldsymbol{g}) - Q_p(\bar{\boldsymbol{g}}_n^*) &= \{l_p(\Phi\boldsymbol{\beta}) - l_p(\Phi\bar{\boldsymbol{\beta}}_n^*)\} \\ &\quad + \Big\{ -\sum_{j=1}^{p} p_{\lambda_n}(\|(\Phi\boldsymbol{\beta})_j\|_{L_2}) + \sum_{j=1}^{p} p_{\lambda_n}(\|(\Phi\bar{\boldsymbol{\beta}}_n^*)_j\|_{L_2})\Big\} \\ &= J_3(\boldsymbol{\beta}) + J_4(\boldsymbol{\beta}), \end{aligned} \tag{35}$$

where $(\Phi\boldsymbol{\beta})_j$ and $(\Phi\bar{\boldsymbol{\beta}}_n^*)_j$ are the jth element of $\Phi\boldsymbol{\beta}$ and $\Phi\bar{\boldsymbol{\beta}}_n^*$.

We evaluate $J_3(\boldsymbol{\beta})$ on $\bar{\Gamma}_M$ as in the proof of Theorem 1.

$$\begin{aligned} J_3(\boldsymbol{\beta}) &= (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_n^*)^\top S_p(\bar{\boldsymbol{\beta}}_n^*) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_n^*)^\top D_p(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_n^*) \\ &= J_{31}(\boldsymbol{\beta}) + J_{32}(\boldsymbol{\beta}), \end{aligned} \tag{36}$$

where $\bar{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}_n^*$. By applying Proposition 2 and Proposition 4 with $\boldsymbol{\beta}_1 = \bar{\boldsymbol{\beta}}_n^*$, we obtain on $\bar{\Gamma}_M$,

$$J_{31}(\boldsymbol{\beta}) = Mr_{pn}\mathcal{O}_p(r_{pn}) + \mathcal{O}_p(Mr_{pn}\rho_n) = \mathcal{O}_p(Mr_{pn}^2). \tag{37}$$

Proposition 3 implies that there is a positive constant M_2 such that

$$J_{32}(\boldsymbol{\beta}) \leq -M_2M^2r_{pn}^2 \tag{38}$$

uniformly on $\bar{\Gamma}_M$ with probability tending to 1. Thus (37) and (38) yield that there is a positive constant M such that

$$\lim_{M\to\infty}\lim_{n\to\infty} \mathsf{P}\Big(\sup_{\Phi\boldsymbol{\beta}\in\bar{\Gamma}_M} J_3(\boldsymbol{\beta}) \leq -M_2M^2r_{pn}^2/2\Big) = 1 \tag{39}$$

Next we deal with $J_4(\boldsymbol{\beta})$. Here recall that $\bar{g}_{nj}^* = (\Phi\bar{\boldsymbol{\beta}}_n^*)_j = 0,\ j = s+1,\ldots,p$, $\min_{1\leq j\leq s}\|g_{0j}\|_{L_2}/\lambda_n \to \infty$, and $\|\bar{g}_{nj}^* - g_{0j}\|_{L_2} = \mathcal{O}(\rho_n) = o(\lambda_n),\ j = 1,\ldots,p$.

By Lemma 1 and the above facts, we have on $\bar{\Gamma}_M$,

$$\|(\Phi\boldsymbol{\beta})_j\|_{L_2} > a_0\lambda_n \quad \text{and} \quad \|(\Phi\bar{\boldsymbol{\beta}}_n^*)_j\|_{L_2} > a_0\lambda_n,\ j = 1,\ldots,s, \tag{40}$$

and

$$\|(\Phi\boldsymbol{\beta})_j\|_{L_2} \leq Cr_{pn} = o(\lambda_n),\ j = s+1,\ldots,p. \tag{41}$$

We obtain by Assumption P, (40), and (41) that

$$J_4(\boldsymbol{\beta}) = -\sum_{j=s+1}^{p} p_{\lambda_n}(\|(\Phi\boldsymbol{\beta})_j\|_{L_2}) \leq 0 \quad \text{on } \bar{\Gamma}_M. \tag{42}$$

(34) follows from (35), (39), and (42). Hence the proof of Theorem 2 is complete.

Proof of Theorem 3. Let $\hat{\boldsymbol{g}}_n = (\hat{g}_{n1},\ldots,\hat{g}_{np})^\top$ be a local maximizer of $Q_p(\boldsymbol{g})$ satisfying the condition of Theorem 3. First we establish the sparsity (i).

We choose l from $\{s+1,\ldots,p\}$. If $\hat{g}_{nl} \neq 0$, we replace $\hat{g}_{nl}$ of $\hat{\boldsymbol{g}}_n$ with 0 and denote it by $\hat{\boldsymbol{g}}_{nl}$. We also define $\hat{\boldsymbol{g}}_t$, $0 \leq t \leq 1$, by

$$\hat{\boldsymbol{g}}_t = \hat{\boldsymbol{g}}_n + t(\hat{\boldsymbol{g}}_{nl} - \hat{\boldsymbol{g}}_n) = (1-t)\hat{\boldsymbol{g}}_n + t\hat{\boldsymbol{g}}_{ln} \tag{43}$$

and compare $Q_p(\hat{\boldsymbol{g}}_t)$ and $Q_p(\hat{\boldsymbol{g}}_n)$. Conditions on $\hat{\boldsymbol{g}}_n$ imply that $\|\hat{\boldsymbol{g}}_n\|_\infty \leq \|\boldsymbol{g}_0\|_\infty + A_n d_n r_{pn} = \mathcal{O}(1)$ and $0 < \|\hat{g}_{nl}\| \leq d_n r_{pn} = o(\lambda_n)$. Note that $\|\hat{\boldsymbol{g}}_{nl} - \hat{\boldsymbol{g}}_n\|_{L_2} = \|\hat{g}_{nl}\|_{L_2}$.

We represent $Q_p(\hat{\boldsymbol{g}}_t) - Q_p(\hat{\boldsymbol{g}}_n)$ as

$$\begin{aligned} & Q_p(\hat{\boldsymbol{g}}_t) - Q_p(\hat{\boldsymbol{g}}_n) \\ &= \{l_p(\hat{\boldsymbol{g}}_t) - l_p(\hat{\boldsymbol{g}}_n)\} + \{-p_{\lambda_n}((1-t)\|\hat{g}_{nl}\|_{L_2}) + p_{\lambda_n}(\|\hat{g}_{nl}\|_{L_2})\} \\ &= J_5 + J_6. \end{aligned} \tag{44}$$

It is easy to see that for some $\bar{t} \in [0, t]$,

$$J_6 = t\|\hat{g}_{nl}\|_{L_2} p'_{\lambda_n}((1-\bar{t})\|\hat{g}_{nl}\|_{L_2}). \tag{45}$$

We evaluate J_5 by employing Propositions 2-4. By using the orthonormal basis, we can represent $\hat{\boldsymbol{g}}_n$ and $\hat{\boldsymbol{g}}_t$ as $\hat{\boldsymbol{g}}_n = \Phi\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{g}}_t = \Phi\hat{\boldsymbol{\beta}}_t$ for some $\hat{\boldsymbol{\beta}}_n \in R^{pK_n}$ and $\hat{\boldsymbol{\beta}}_t \in R^{pK_n}$, respectively. Then we have

$$\begin{aligned} J_5 &= l_p(\Phi\hat{\boldsymbol{\beta}}_t) - l_p(\Phi\hat{\boldsymbol{\beta}}_n) \\ &= (\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_n)^\top S_p(\hat{\boldsymbol{\beta}}_n) + (\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_n)^\top D_p(\bar{\boldsymbol{\beta}}_t)(\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_n) \\ &= J_{51} + J_{52}, \end{aligned} \tag{46}$$

where $\bar{\boldsymbol{\beta}}_t$ is between $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_n$.

By Propositions 2 and 4, we have uniformly in t,

$$J_{51} = t\|\hat{\boldsymbol{g}}_{nl} - \hat{\boldsymbol{g}}_n\|\{\mathcal{O}_p(r_{pn}) + \mathcal{O}_p(d_n r_{pn})\} = t\|\hat{g}_{nl}\|_{L_2}\mathcal{O}_p(d_n r_{pn}). \tag{47}$$

By Proposition 3, we have uniformly in t,

$$J_{52} = t^2\|\hat{\boldsymbol{g}}_{nl} - \hat{\boldsymbol{g}}_n\|^2\mathcal{O}_p(1) = t\|\hat{g}_{nl}\|_{L_2}\mathcal{O}_p(d_n r_{pn}). \tag{48}$$

(47) and (48) imply that uniformly in t,

$$J_5 = t\|\hat{g}_{nl}\|_{L_2}\mathcal{O}_p(d_n r_{pn}). \tag{49}$$

Note that $\mathcal{O}_p(d_n r_{pn})$ in (47) and (48) are independent of any particular choice of $\hat{\boldsymbol{g}}_n$.

By combining (45), (49), and Assumption P, we have uniformly in $t \in (0, 1/2)$,

$$Q_p(\hat{\boldsymbol{g}}_t) - Q_p(\hat{\boldsymbol{g}}_n) = t\|\hat{g}_{nl}\|_{L_2}\{\mathcal{O}_p(d_n r_{pn}) + p'_{\lambda_n}((1-\bar{t})\|\hat{g}_{nl}\|_{L_2})\} > 0 \tag{50}$$

with probability tending to 1. This contradicts the local optimality of $\hat{\boldsymbol{g}}_n$. Hence we have $\hat{g}_{nl} = 0$, $l = s+1, \ldots, p$, for any local maximizer $\hat{\boldsymbol{g}}_n$ in Theorem 3 with probability tending to 1. Hence the proof of the latter half is complete.

Next we prove that the local maximizer in Theorem 3 has the oracle property. Let $\hat{\boldsymbol{g}}_n$ be the local maximizer in Theorem 3 such that $\hat{g}_{nj} = 0, j = s+1, \ldots, p$. Then we consider $\hat{\boldsymbol{g}}_n + \boldsymbol{\eta}$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top \in \mathbf{G}_0$ and $\eta_j = 0$, $j = s+1, \ldots, p$. By Assumption P and the local optimality of $\hat{\boldsymbol{g}}_n$, there is a small positive constant ϵ such that

$$\sum_{j=1}^{s} p_{\lambda_n}(\|\hat{g}_{nj} + \eta_j\|_{L_2}) = \sum_{j=1}^{s} p_{\lambda_n}(\|\hat{g}_{nj}\|_{L_2}) \text{ and } Q_p(\hat{\boldsymbol{g}}_n + \boldsymbol{\eta}) \leq Q_p(\hat{\boldsymbol{g}}_n) \quad (51)$$

when $\|\boldsymbol{\eta}\|_{L_2} \leq \epsilon$. Since $l_s(\boldsymbol{g}) = l_p(\boldsymbol{g})$ with $g_j = 0$, $j = s+1, \ldots, p$, we have by (51) that

$$l_s(\hat{\boldsymbol{g}}_n + \boldsymbol{\eta}) = l_p(\hat{\boldsymbol{g}}_n + \boldsymbol{\eta}) \leq l_p(\hat{\boldsymbol{g}}_n) = l_s(\hat{\boldsymbol{g}}_n). \quad (52)$$

(52) means that $\hat{\boldsymbol{g}}_n$ is a local maximizer of $l_s(\boldsymbol{g})$. Since the assumptions in Theorem 1 are satisfied with p replaced with s, $\hat{\boldsymbol{g}}_n$ is the unique maximizer of $l_s(\boldsymbol{g})$ with probability tending to 1 and has the desired convergence rate. Hence the latter half of Theorem 3 is established.

Proof of Theorem 4. First as in the proof of Theorem 2, we demonstrate that for any positive ϵ, there is a positive constant M such that

$$\lim_{n\to\infty} \mathsf{P}\Big(\sup_{\boldsymbol{g}\in\bar{\Gamma}_M} \bar{Q}_p(\boldsymbol{g}) < \bar{Q}_p(\bar{\boldsymbol{g}}_n^*)\Big) > 1 - \epsilon. \quad (53)$$

Write

$$\begin{aligned} \bar{Q}_p(\boldsymbol{g}) - \bar{Q}_p(\bar{\boldsymbol{g}}_n^*) &= \{l_p(\Phi\boldsymbol{\beta}) - l_p(\Phi\bar{\boldsymbol{\beta}}_n^*)\} \quad (54) \\ &\quad + \Big\{ -\sum_{j=1}^{p} \lambda_n' w_j \|(\Phi\boldsymbol{\beta})_j\|_{L_2} + \sum_{j=1}^{p} \lambda_n' w_j \|(\Phi\bar{\boldsymbol{\beta}}_n^*)_j\|_{L_2} \Big\} \\ &= J_3(\boldsymbol{\beta}) + J_7(\boldsymbol{\beta}). \end{aligned}$$

We evaluated $J_3(\boldsymbol{\beta})$ on $\bar{\Gamma}_M$ as in the proof of Theorem 2. See (39).

We evaluate $J_7(\boldsymbol{\beta})$ on $\bar{\Gamma}_M$ and obtain

$$J_7(\boldsymbol{\beta}) \leq \sum_{j=1}^{s} \lambda_n' w_j \|(\Phi(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_n^*))_j\|_{L_2} \leq \frac{\lambda_n' M\sqrt{s} r_{pn}^2 \max_{1\leq j\leq s} w_j}{r_{pn}}. \quad (55)$$

(53) follows from (39) and (55). (53) implies that a local maximizer $\bar{\boldsymbol{g}}_n$ of $\bar{Q}_p(\boldsymbol{g})$ exists inside $\bar{\Gamma}_M$ with probability tending to 1. Besides, $\bar{Q}_p(\Phi\boldsymbol{\beta})$ is strictly concave inside $\bar{\Gamma}_M$ with probability tending to 1 by Proposition 3. Therefore the local maximizer must be a unique maximizer with probability tending to 1. Hence the proof of Theorem 4 is complete.

Proof of Theorem 5. We proceed as in the proof of Theorem 3. We define $\hat{\boldsymbol{g}}_n$, $\hat{\boldsymbol{g}}_{nl}$, $\hat{\boldsymbol{g}}_t$, $\hat{\boldsymbol{\beta}}_n$, and $\hat{\boldsymbol{\beta}}_t$ as in the proof of Theorem 3. Then choosing l as in the proof of Theorem 3, we have

$$\begin{aligned}
&\bar{Q}_p(\hat{\boldsymbol{g}}_t) - \bar{Q}_p(\hat{\boldsymbol{g}}_n) \\
&= \{l_p(\hat{\boldsymbol{g}}_t) - l_p(\hat{\boldsymbol{g}}_n)\} + \lambda'_n w_l t \|\hat{g}_{nl}\|_{L_2} \\
&= t\|\hat{g}_{nl}\|_{L_2}\mathcal{O}_p(r_{pn}) + \lambda'_n w_l t\|\hat{g}_{nl}\|_{L_2} \\
&= t\|\hat{g}_{nl}\|_{L_2} r_{pn}\Big\{\mathcal{O}_p(1) + \frac{\lambda'_n w_l}{r_{pn}}\Big\} > 0
\end{aligned}$$

uniformly in $t \in (0, 1/2)$ with probability tending to 1. The above inequality contradicts the optimality of $\hat{\boldsymbol{g}}_n$. Hence $\|\hat{g}_{nl}\|_{L_2} = 0$, $l = s+1, \ldots, p$ with probability tending to 1. The sparsity and the optimality of $\hat{\boldsymbol{g}}_n$ implies that $\hat{\boldsymbol{g}}_n$ is equal to the unique maximizer of $\bar{Q}_s(\boldsymbol{g})$. Since the assumptions in Theorem 4 are satisfied with p replaced with s, we have the desired L_2 convergence rate. Hence the proof of Theorem 5 is complete.

6. Coefficients varying with $U(t)$

In this section, we consider another Cox regression model with varying coefficients. We can establish almost the same theoretical results as in section 3 by using the results in [17]. The proofs are similar to those of Theorems 1-5 and we present only the assumptions, the theorems, and a remark on the proofs.

We observe two kinds of covariates, $\boldsymbol{X}(t)$ and $U(t)$, and $U(t)$ is an index variable. It is reasonable to assume that the coefficients of the p-dimensional covariate $\boldsymbol{X}(t)$ are functions of $U(t)$. Specifically, we observe another important influential covariate $U(t)$ in addition to $(Y, \delta, \boldsymbol{X}(t))$ and the hazard function of the failure time T is given by

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{g}_0^\top(U_i(t))\boldsymbol{X}_i(t)\} \tag{56}$$

instead of (1). We assume that $U(t)$ is one-dimensional for simplicity of presentation. In this setup, suppose that we have n i.i.d. observations $(Y_i, \delta_i, \boldsymbol{X}_i(t), U_i(t))$ on $[0, \tau]$, where $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^\top$, and carry out variable selection and estimation of $\boldsymbol{g}_0$ simultaneously. In this section, we take $X_{i1}(t) = 1$ on $[0, \tau]$ and always include $X_{i1}(t)$ in the model.

We describe the assumptions and the norms before we present the theoretical results. We assumed Assumptions A1-3 for the time-varying coefficient model. We should just replace $\boldsymbol{X}_i(t)$ and $\boldsymbol{g}_0(t)$ with $(\boldsymbol{X}_i(t), U_i(t))$ and $\boldsymbol{g}_0(U_i(t))$ or $\boldsymbol{g}_0(u)$, respectively in those assumptions. We call them Assumptions A1'-3' in this section. This model and related ones are considered in several papers, for example, [12], [8], and [10]. They employed local polynomial regression to estimate the nonparametric components with p fixed and examined the asymptotics.

We use almost the same notation and assumptions as in the time-varying coefficient case. However, some conformable changes below are in order. We set $\Omega(U(t), t) = \mathsf{E}\{\boldsymbol{X}(t)\boldsymbol{X}^\top(t) \,|\, U(t)\}$ and denote the density function of $U(t)$ by $f_t(u)$.

Assumption X':
(i) There are positive constants C'_m and C'_M such that $C'_m \leq \lambda_{\min}(\Omega(U(t), t)) \leq \lambda_{\max}(\Omega(U(t), t)) \leq C'_M$ uniformly in t a.s.
(ii) $X_{ij}(t)$ is uniformly bounded in i, j, and t a.s.
(iii) The support of $U(t)$ is $[0, 1]$ and there are positive constants C_L and C_U such that $C_l < f_t(u) < C_U$ uniformly in t.

Assumption M':
(i) There is a positive constant C'_Z such that $\mathsf{E}\{Z(t)|\boldsymbol{X}(t), U(t)\} \geq C'_Z$ uniformly in t a.s.
(ii) There are positive constants C_{L1} and C_{L2} such that $C_{L1} \leq \lambda_0(t) \leq C_{L2}$ on $[0, \tau]$.

In Assumptions G and B, t and τ should be replaced with u and 1, respectively and we call them Assumptions G' and B', respectively. We add an identifiability constraint to Assumption G' later in this section.

We define $\mathbf{H}_0$, $\mathbf{G}_0$, $\|\cdot\|_\infty$, and $\|\cdot\|_{L_2}$ almost in the same way as in section 2. For example, for $\boldsymbol{g}(u) = (g_1(u), \ldots, g_p(u))^\top$ on $[0, 1]$,

$$\|\boldsymbol{g}\|_{L_2}^2 = \sum_{j=1}^{p} \|g_j\|_{L_2}^2 = \sum_{j=1}^{p} \int_0^1 g_j^2(u)du. \tag{57}$$

We can define ρ_n, r_{pn}, r_{sn}, and Assumption RA in the same way as in section 2.

It is crucial to introduce suitable norms and an suitable identifiability constraint on $\mathbf{H}_0$ and $\mathbf{G}_0$ when we employ the results of [17]. One reason is that we have $l_p(\boldsymbol{h}) = l_p(\boldsymbol{h}+c\boldsymbol{e}_1)$ in this model, where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top \in R^p$ and $l_p(\boldsymbol{h})$ is the partial likelihood defined in (64) below. Therefore we adapt $\|\cdot\|$ and $\|\cdot\|_0$ to this setup and impose an identifiability constraint as in [17].

For this model, we define $\|\boldsymbol{h}\|$ on $\mathbf{H}_0$ in terms of the inner product defined in (58) below.

$$(\boldsymbol{h}_1, \boldsymbol{h}_2) = \int_0^\tau (\boldsymbol{h}_1, \boldsymbol{h}_2)_Z(t) d\,\mathsf{E}\{N(t)\}, \tag{58}$$

where

$$(\boldsymbol{h}_1, \boldsymbol{h}_2)_Z(t) = \mathsf{E}[\{\boldsymbol{h}_1^\top(U(t))\boldsymbol{X}(t)\}\{\boldsymbol{h}_2^\top(U(t))\boldsymbol{X}(t)\}Z(t)]/\,\mathsf{E}\{Z(t)\}.$$

The identifiability constraint is imposed through the equation

$$\int_0^\tau \frac{\mathsf{E}\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)Z(t)\}}{\mathsf{E}\{Z(t)\}} d\,\mathsf{E}\{N(t)\} = 0. \tag{59}$$

As in [17], we define $\mathbf{H}$ by

$$\mathbf{H} = \{\boldsymbol{h} \in \mathbf{H}_0 |\, \boldsymbol{h} \text{ satisfies (59).}\}.$$

For any $\boldsymbol{h} \in \mathbf{H}_0$, we have

$$\boldsymbol{h} - (m, 0, \ldots, 0)^\top \in \mathbf{H}, \tag{60}$$

where

$$m = [\mathsf{E}\{Z(\tau)\}]^{-1} \int_0^\tau \frac{\mathsf{E}\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)Z(t)\}}{\mathsf{E}\{Z(t)\}} d\,\mathsf{E}\{N(t)\},$$

and any element of $\mathbf{H}$ can be written as in (60). This means the constraint (59) affects only the first element of $\boldsymbol{h}$. Hereafter we assume that $\boldsymbol{g}_0 \in \mathbf{H}$ and include this into Assumption G'. We define $\|\boldsymbol{h}\|_0$ on $\mathbf{H}$ in terms of the inner product in (61) below.

$$\begin{aligned}
&(\boldsymbol{h}_1, \boldsymbol{h}_2)_0 \qquad\qquad (61)\\
&= \int_0^\tau [\mathsf{E}\{Z(t)\}]^{-1}\, \mathsf{E}([\boldsymbol{h}_1^\top(U(t))\boldsymbol{X}(t) - \mathsf{E}\{\boldsymbol{h}_1^\top(U(t))\boldsymbol{X}(t)\}/\,\mathsf{E}\{Z(t)\}]\\
&\qquad \times[\boldsymbol{h}_2^\top(U(t))\boldsymbol{X}(t) - \mathsf{E}\{\boldsymbol{h}_2^\top(U(t))\boldsymbol{X}(t)\}/\,\mathsf{E}\{Z(t)\}]Z(t))d\,\mathsf{E}\{N(t)\}.
\end{aligned}$$

We also define the empirical versions of $\|\boldsymbol{h}\|$ and $\|\boldsymbol{h}\|_0$ by replacing $\mathsf{E}(\cdot)$ with the empirical measure $\mathsf{E}_n(\cdot)$ in the definitions. We denote the empirical versions by $\|\boldsymbol{h}\|_n$ and $\|\boldsymbol{h}\|_{0n}$ as in [17] and the time-varying case. The empirical version of the constraint (59) is also given by replacing $\mathsf{E}(\cdot)$ with the empirical measure $\mathsf{E}_n(\cdot)$ in (59). Then we define $\mathbf{G}$ by

$$\mathbf{G} = \{\boldsymbol{g} \in \mathbf{G}_0 | \, \boldsymbol{g} \text{ satisfies the empirical version of (59).}\}.$$

For any $\boldsymbol{g} \in \mathbf{G}_0$, we have

$$\boldsymbol{g} - (m_n, 0, \ldots, 0)^\top \in \mathbf{G}, \tag{62}$$

where

$$m_n = [\mathsf{E}_n\{Z(\tau)\}]^{-1} \int_0^\tau \frac{\mathsf{E}_n\{\boldsymbol{g}^\top(U(t))\boldsymbol{X}(t)Z(t)\}}{\mathsf{E}_n\{Z(t)\}} d\mathsf{E}_n\{N(t)\}.$$

On the other hand, any element of $\mathbf{G}$ can be written as in (62). This empirical identifiability constraint also affects only the first element of $\boldsymbol{g}$. Note that the dimension of $\mathbf{G}_0$ is pK_n and that of $\mathbf{G}$ is $pK_n - 1$.

The equivalence between $\|\boldsymbol{h}\|$ and $\|\boldsymbol{h}\|_{L_2}$ on $\mathbf{H}_0$ and that between $\|\boldsymbol{h}\|_0$ and $\|\boldsymbol{h}\|_{L_2}$ on $\mathbf{H}$ are crucial to the derivation of the main results as in [17] and we state them in Lemmas 5-6. If only those equivalences are established, we can proceed as in the time-varying case and [17] with just conformable changes. The proofs of Lemmas 5-8 are postponed to section 7.

Lemma 5. *Suppose that Assumptions A1'-3', X', M', B', and G' hold. Then uniformly in* $\boldsymbol{h} \in \mathbf{H}_0$,

$$\|\boldsymbol{h}\| \sim \|\boldsymbol{h}\|_{L_2}.$$

Lemma 6. *Suppose that Assumptions A1'-3', X', M', B', and G' hold. Then uniformly in* $\boldsymbol{h} \in \mathbf{H}$,

$$\|\boldsymbol{h}\|_0 \sim \|\boldsymbol{h}\|_{L_2}.$$

Next we consider the empirical versions.

Lemma 7. *Suppose that Assumptions A1'-3', X', M', B', G', and RA hold. Then*

$$\sup_{\boldsymbol{g}_1, \boldsymbol{g}_2 \in \mathbf{G}_0} \left| \frac{(\boldsymbol{g}_1, \boldsymbol{g}_2)_n - (\boldsymbol{g}_1, \boldsymbol{g}_2)}{\|\boldsymbol{g}_1\| \|\boldsymbol{g}_2\|} \right| = o_p(1).$$

In [17], another function space $\tilde{\mathbf{G}}$ is introduced for technical reasons and we follow them. As we comment in Remark 3 below, the difference between $\mathbf{G}$ and $\tilde{\mathbf{G}}$ defined in (63) below does not affect the proofs of Theorems 6-10.

We define the theoretical version $\tilde{\mathbf{G}}$ of $\mathbf{G}$ by

$$\tilde{\mathbf{G}} = \{\boldsymbol{g} \in \mathbf{G}_0 | \, \boldsymbol{g} \text{ satisfies (59).}\} \tag{63}$$

Lemma 8. *Suppose that Assumptions A1'-3', X', M', B', G', and RA hold. Then*

$$\sup_{\boldsymbol{g}_1, \boldsymbol{g}_2 \in \tilde{\mathbf{G}}} \Big| \frac{(\boldsymbol{g}_1, \boldsymbol{g}_2)_{0n} - (\boldsymbol{g}_1, \boldsymbol{g}_2)_0}{\|\boldsymbol{g}_1\|_0 \|\boldsymbol{g}_2\|_0} \Big| = o_p(1).$$

Now we define the partial likelihood $l_p(\boldsymbol{h})$ and the expected value version $\Lambda_p(\boldsymbol{h})$ for $\boldsymbol{h} \in \mathbf{G}$ or $\mathbf{H}$ as in the time-varying case and [17].

$$\begin{aligned} l_p(\boldsymbol{h}) &= \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \boldsymbol{h}^\top(U_i(t)) \boldsymbol{X}_i(t) dN_i(t) \\ &\quad - \int_0^\tau \log \Big[n^{-1} \sum_{i=1}^{n} Z_i(t) \exp \Big\{ \boldsymbol{h}^\top(U_i(t)) \boldsymbol{X}_i(t) \Big\} \Big] d\bar{N}(t) \end{aligned} \tag{64}$$

and

$$\begin{aligned} \Lambda_p(\boldsymbol{h}) &= \mathsf{E} \Big\{ \int_0^\tau \boldsymbol{h}^\top(U(t)) \boldsymbol{X}(t) dN(t) \Big\} \\ &\quad - \int_0^\tau \log \Big(\mathsf{E} \Big[Z(t) \exp \Big\{ \boldsymbol{h}^\top(U(t)) \boldsymbol{X}(t) \Big\} \Big] \Big) d\, \mathsf{E}\{N(t)\}. \end{aligned} \tag{65}$$

To carry out variable selection and estimation of $\boldsymbol{g}_0$ simultaneously, we define $Q_p(\boldsymbol{g})$ and $\bar{Q}_p(\boldsymbol{g})$ for $\boldsymbol{g} \in \mathbf{G}$ as in section 2. The former gives the group SCAD-type estimator and the latter gives the adaptive group Lasso estimator.

$$Q_p(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \sum_{j=2}^{p} p_{\lambda_n}(\|g_j\|_{L_2}) \tag{66}$$

$$\bar{Q}_p(\boldsymbol{g}) = l_p(\boldsymbol{g}) - \lambda_n' \sum_{j=2}^{p} w_j \|g_j\|_{L_2} \tag{67}$$

We state the properties of the maximum partial likelihood estimator $\tilde{\boldsymbol{g}}_n$, the group SCAD-type estimator $\hat{\boldsymbol{g}}_n$, and the adaptive group Lasso estimator $\bar{\boldsymbol{g}}_n$. We define them as in (14) and (19) by replacing $\mathbf{G}_0$ with $\mathbf{G}$. Especially we are interested in the sparsity and oracle property of $\hat{\boldsymbol{g}}_n$ and $\bar{\boldsymbol{g}}_n$ under Assumption S. We can also define $l_s(\boldsymbol{g})$, $Q_s(\boldsymbol{g})$, and $\bar{Q}_s(\boldsymbol{g})$ by taking $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{is}(t))^\top$ and $\boldsymbol{g} = (g_1, \ldots, g_s)^\top$ as in section 2.

Now we state the main theoretical results of this section. Recall that $\boldsymbol{g}_0 \in \mathbf{H}$ in this paper. The L_2 convergence rate of the maximum partial likelihood estimator is given in Theorem 6. The SCAD-type estimator and the adaptive group Lasso estimator are considered in Theorems 7-8 and Theorems 9-10, respectively.

Theorem 6. *Suppose that Assumptions A1'-3', X', M', G', B', and RA hold and $r_{pn} \to 0$. Then with probability tending to 1, there is a unique maximizer $\tilde{\boldsymbol{g}}_n = (\tilde{g}_{n1}, \ldots, \tilde{g}_{np})^\top$ of $l_p(\boldsymbol{g})$ over $\mathbf{G}$ and we have*

$$\|\tilde{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} = \mathcal{O}_p(r_{pn}).$$

Theorem 7. *Suppose that all the assumptions in Theorem 6 and Assumptions P and S hold. Then for any positive ϵ, there is a positive constant M such that*

$$\lim_{n\to\infty} \mathsf{P}(\text{ There is a local maximizer } \hat{\boldsymbol{g}}_n \text{ of } Q_p(\boldsymbol{g}) \text{ over } \mathbf{G} \text{ such that } \|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} \le M r_{pn}) > 1 - \epsilon.$$

Theorem 8. *Suppose that the assumptions in Theorem 7 hold and let $\{d_n\}$ be a sequence of positive numbers satisfying $d_n \to \infty$, $\lambda_n/(d_n r_{pn}) \to \infty$, and $A_n d_n r_{pn} = \mathcal{O}(1)$.*
(i) *With probability tending to 1, any local maximizer $\hat{\boldsymbol{g}}_n = (\hat{g}_{n1}, \ldots, \hat{g}_{np})^\top$ of $Q_p(\boldsymbol{g})$ over $\mathbf{G}$ satisfying $\|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} \le d_n r_{pn}$ has the sparsity.*
(ii) *With probability tending to 1, the local maximizer in* (i) *is the unique maximizer of $l_s(\boldsymbol{g})$ and satisfies*

$$\sum_{j=1}^{s} \|\hat{g}_{nj} - g_{0j}\|_{L_2}^2 = \mathcal{O}(r_{sn}^2).$$

Theorem 9. *Suppose that the assumptions in Theorem 6 and Assumption S hold and that $\lambda_n' \sqrt{s} \max_{2\le j\le s} w_j / r_{pn} = \mathcal{O}_p(1)$. Then with probability tending*

to 1, there is a unique maximizer $\bar{\boldsymbol{g}}_n = (\bar{g}_{n1}, \ldots, \bar{g}_{np})^\top$ *of* $\bar{Q}_p(\boldsymbol{g})$ *over* $\mathbf{G}$ *and we have*

$$\|\bar{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_{L_2} = \mathcal{O}_p(r_{pn}).$$

Remark 3. We consider the weights $\{w_j\}$ in this remark. Suppose we take $w_j = 1/\|\tilde{g}_{nj}\|_{L_2}$. If we have $\lambda'_n\sqrt{s}/(r_{pn}\min_{2\le j\le s}\|g_{0j}\|_{L_2}) = \mathcal{O}(1)$ and $\min_{2\le j\le s}\|g_{0j}\|_{L_2}/r_{pn} \to \infty$, the assumption on $\{w_j\}$ in Theorem 9 is satisfied. If we also have $\lambda'_n/r_{pn}^2 \to \infty$, the assumptions on $\{w_j\}$ in Theorem 10 below are also satisfied.

Theorem 10. *Suppose that the assumptions in Theorem 9 hold and that* $\lambda'_n \min_{s<j\le p} w_j/(r_{pn}) \to \infty$ *in probability and* $A_n r_{pn} = \mathcal{O}(1)$. *Then with probability tending to 1, the unique maximizer* $\bar{\boldsymbol{g}}_n$ *has the sparsity and is equal to the unique maximizer of* $\bar{Q}_s(\boldsymbol{g})$. *In addition we have*

$$\sum_{j=1}^{s} \|\bar{g}_{nj} - g_{0j}\|_{L_2}^2 = \mathcal{O}_p(r_{sn}^2).$$

We give two remarks here. One is about the proofs of Theorems 6-10 and the other is about computation of the estimators.

Remark 4. In [17], the authors maximize $l_p(\boldsymbol{g})$ over $\tilde{\mathbf{G}}$, not $\mathbf{G}$, in the proofs and they examined the asymptotics of the maximizer over $\tilde{\mathbf{G}}$ closely. And then they proved that the difference between the maximizer over $\tilde{\mathbf{G}}$ and that over $\mathbf{G}$ is $\mathcal{O}_p(r_{pn})$. This method also works well in the setup of this paper because the identifiability constraint affects only the first element of $\boldsymbol{g}$, g_1, and the penalty term does not include the first element. Besides, when we obtain the maximizer over $\mathbf{G}$ from that over $\tilde{\mathbf{G}}$ as in (62), the difference m_n does not have any influences on the L_2 convergence rate as in the proof of Lemma 9 of [17].

Remark 5. We can use any basis and identifiability constraint on the first element of $\boldsymbol{g}$, g_1, when we compute $\tilde{\boldsymbol{g}}_n$, $\hat{\boldsymbol{g}}_n$, and $\bar{\boldsymbol{g}}_n$. An orthonormal basis is chosen in the proofs only for technical reasons and the partial likelihood or $\|g_j\|_{L_2}$ does not depend on any particular choice of basis. Besides, identifiability constraints on g_1 does not affect g_j, $j = 2, \ldots, p$. Therefore we can use a B-spline basis and an identifiability constraint $g_1(0) = 0$ when we compute $\tilde{\boldsymbol{g}}_n$, $\hat{\boldsymbol{g}}_n$, and $\bar{\boldsymbol{g}}_n$.

7. Technical proofs

Lemmas 1-8 and Propositions 1-4 are proved in this section.

PROOF OF LEMMA 1. It is easy to see that

$$\mathsf{E}[\{\boldsymbol{X}^\top(t)\boldsymbol{h}(t)\}^2 Z(t)] = \mathsf{E}[\{\boldsymbol{X}^\top(t)\boldsymbol{h}(t)\}^2 \mathsf{E}\{Z(t) \mid \boldsymbol{X}(t)\}] \sim \boldsymbol{h}^\top(t)\Omega(t)\boldsymbol{h}(t)$$

uniformly in t and $\boldsymbol{h}$ due to Assumptions X and M.

Thus we have uniformly in $\boldsymbol{h}$,

$$\begin{aligned} \|\boldsymbol{h}\|^2 &= \int_0^\tau \frac{\mathsf{E}[\{\boldsymbol{X}^\top(t)\boldsymbol{h}(t)\}^2 Z(t)]}{\mathsf{E}\{Z(t)\}} d\,\mathsf{E}\{N(t)\} \\ &\sim \int_0^\tau |\boldsymbol{h}(t)|^2\, \mathsf{E}[\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}]\lambda_0(t)dt \\ &\sim \|\boldsymbol{h}\|_{L_2}^2. \end{aligned}$$

Note that $|\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)|$ is uniformly bounded in t by Assumptions M and G. Hence the proof of Lemma 1 is complete.

PROOF OF LEMMA 2. We have uniformly in $\boldsymbol{h}$,

$$\begin{aligned} \|\boldsymbol{h}\|_0^2 &= \int_0^\tau [\mathsf{E}\{Z(t)\}]^{-1}\, \mathsf{E}([\boldsymbol{h}^\top(t)\boldsymbol{X}(t) - \mathsf{E}\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t)Z(t)\}/\,\mathsf{E}\{Z(t)\}]^2 \\ &\qquad \times Z(t))d\,\mathsf{E}\{N(t)\} \\ &= \int_0^\tau \inf_c([\mathsf{E}\{Z(t)\}]^{-1}\, \mathsf{E}[\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t) - c\}^2 Z(t)])d\,\mathsf{E}\{N(t)\} \\ &\sim \int_0^\tau \inf_c \mathsf{E}[\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t) - c\}^2 Z(t)]d\,\mathsf{E}\{N(t)\}. \end{aligned}$$

We evaluate the integrand and obtain

$$\begin{aligned} &\inf_c \mathsf{E}[\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t) - c\}^2 Z(t)] \\ &\quad = \inf_c \mathsf{E}[\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t) - c\}^2\, \mathsf{E}\{Z(t)|\boldsymbol{X}(t)\}] \\ &\quad \sim \inf_c \mathsf{E}[\{\boldsymbol{h}^\top(t)\boldsymbol{X}(t) - c\}^2] \\ &\quad = \boldsymbol{h}^\top(t)\Sigma(t)\boldsymbol{h}(t) \sim |\boldsymbol{h}(t)|^2. \end{aligned}$$

uniformly in $\boldsymbol{h}$ due to Assumptions M and X. Thus we have uniformly in $\boldsymbol{h}$,

$$\|\boldsymbol{h}\|_0^2 \sim \int_0^\tau |\boldsymbol{h}(t)|^2\, \mathsf{E}[\exp\{\boldsymbol{g}_0(t)^\top\boldsymbol{X}(t)\}]\lambda_0(t)dt \sim \|\boldsymbol{h}\|_{L_2}^2.$$

Hence the proof of Lemma 2 is complete.

PROOF OF LEMMA 3. We omit the details since we can prove Lemma 3 almost in the same way as Lemma 3 of [17] by just replacing $f_1(\boldsymbol{X}(t))$ and $f_2(\boldsymbol{X}(t))$ there with $\boldsymbol{g}_1^\top(t)\boldsymbol{X}(t)$ and $\boldsymbol{g}_2^\top(t)\boldsymbol{X}(t)$, respectively. We assume that p can increase moderately. However, this does not affect the application of Lemma 10 of [16] since N_n in [17] increases , too. Notice the typo in the definition of I_3 there.

PROOF OF LEMMA 4. We omit the details since we can also prove Lemma 4 almost in the same way as Lemma 4 of [17] by just replacing $f_1(\boldsymbol{X}(t))$, $f_2(\boldsymbol{X}(t))$, and $\phi_i(\boldsymbol{X}(t))$ with $\boldsymbol{g}_1^\top(t)\boldsymbol{X}(t)$, $\boldsymbol{g}_2^\top(t)\boldsymbol{X}(t)$, and $\boldsymbol{\phi}_i^\top(t)\boldsymbol{X}(t)$, respectively. Note that $\tilde{\mathbf{G}}$ is replaced with $\mathbf{G}_0$ in the time varying coefficient case. The authors of [17] used their Lemma 5 in the proof of their Lemma 4. We can verify their Lemma 5 by replacing $\phi_i(\boldsymbol{X}(t))$ and $h_n(\boldsymbol{X}(t))$ with $\boldsymbol{\phi}_i^\top(t)\boldsymbol{X}(t)$ and 1, respectively.

Before we prove Proposition 1, we state Lemma 9, which corresponds to Lemma 8 of [17]. Recall that $\boldsymbol{g}_0$ is the true coefficient function and $\boldsymbol{g}_0 \in \mathbf{H}_0$.

Lemma 9. *Suppose that the same assumptions hold as in Theorem 1. Then for any positive M, there are positive constants M_1 and M_2 such that*

$$-M_1\|\boldsymbol{h}-\boldsymbol{g}_0\|^2 \le \Lambda_p(\boldsymbol{h}) - \Lambda_p(\boldsymbol{g}_0) \le -M_2\|\boldsymbol{h}-\boldsymbol{g}_0\|^2$$

for any $\boldsymbol{h} \in \mathbf{H}_0$ satisfying $\|\boldsymbol{h}\|_\infty \le M$.

PROOF. This lemma can be proved almost in the same way as Lemma 8 of [17]. We should replace α^* with $\boldsymbol{g}_0$ and set $\boldsymbol{h}_u = \boldsymbol{g}_0 + u(\boldsymbol{h}-\boldsymbol{g}_0)$.

Then we have

$$\begin{aligned}
&\frac{d}{du}\Lambda_p(\boldsymbol{h}_u)\Big|_{u=0} \\
&= \mathsf{E}\Big\{\int_0^\tau (\boldsymbol{h}-\boldsymbol{g}_0)^\top(t)\boldsymbol{X}(t)dN(t)\Big\} \\
&\quad - \int_0^\tau \frac{\mathsf{E}[(\boldsymbol{h}-\boldsymbol{g}_0)^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}]}{E[Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}]} \\
&\qquad \times E[Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}]\lambda_0(t)dt \\
&= \mathsf{E}\Big[\int_0^\tau (\boldsymbol{h}-\boldsymbol{g}_0)^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}\lambda_0(t)dt\Big] \\
&\quad - \int_0^\tau \mathsf{E}[(\boldsymbol{h}-\boldsymbol{g}_0)^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}]\lambda_0(t)dt \\
&= 0.
\end{aligned}$$

We used (24) in the above equations. Thus we can proceed as in [17] with conformable changes. The details are omitted.

PROOF OF PROPOSITION 1. As in [17], we can easily prove that $\Lambda_p(\boldsymbol{h})$ is strictly concave by calculating the second derivative of $\Lambda_p(t\boldsymbol{h}_1 + (1-t)\boldsymbol{h}_2)$. The definition of ρ_n implies that there is $\boldsymbol{g}_{an} \in \mathbf{G}_0$ such that $\|\boldsymbol{g}_{an} - \boldsymbol{g}_0\|_\infty \leq 2\rho_n$. Hence $\|\boldsymbol{g}_{an} - \boldsymbol{g}_0\| \leq 2\rho_n$ for any sufficiently large n and we have by Lemma 9,

$$-4M_1\rho_n^2 + \Lambda_p(\boldsymbol{g}_0) \leq \Lambda_p(\boldsymbol{g}_{an}).$$

On the other hand, we have

$$\Lambda_p(\boldsymbol{g}) \leq \Lambda_p(\boldsymbol{g}_0) - M_3^2 M_2 \rho_n^2$$

on $\{\boldsymbol{g} \in \mathbf{G}_0 \,|\, \|\boldsymbol{g} - \boldsymbol{g}_0\| = M_3\rho_n\}$ for $M_3 > 2$. Here we used the fact that we have eventually

$$\{\boldsymbol{g} \in \mathbf{G}_0 \,|\, \|\boldsymbol{g} - \boldsymbol{g}_0\| = M_3\rho_n\} \subset \{\boldsymbol{g} \in \mathbf{G}_0 \,|\, \|\boldsymbol{g}\|_\infty \leq C\}$$

for some fixed C by Assumption RA. If we choose M_3 such that $M_3 > \sqrt{4M_1/M_2}$, we have on $\{\boldsymbol{g} \in \mathbf{G}_0 \,|\, \|\boldsymbol{g} - \boldsymbol{g}_0\| = M_3\rho_n\}$,

$$\Lambda_p(\boldsymbol{g}) < \Lambda_p(\boldsymbol{g}_{an}). \tag{68}$$

The existence of a unique maximizer $\boldsymbol{g}_n^*$ satisfying $\|\boldsymbol{g}_n^* - \boldsymbol{g}_0\| = \mathcal{O}(\rho_n)$ follows from (68) and the concavity of $\Lambda_p(\boldsymbol{h})$. Hence the proof of Proposition 1 is complete.

PROOF OF PROPOSITION 2. We prove this proposition by following the proof of Lemma 10 of [17].

Write the jth element of $S_p(\boldsymbol{\beta}_n^*)$ as

$$\begin{aligned} S_{pj}(\boldsymbol{\beta}_n^*) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)dN_i(t) \\ &\quad - \int_0^\tau \frac{\sum_{i=1}^n \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)Z_i(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}_i(t)\}}{\sum_{i=1}^n Z_i(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}_i(t)\}} d\bar{N}(t). \end{aligned}$$

Because of the optimality of $\boldsymbol{g}_n^* = \Phi\boldsymbol{\beta}_n^*$, we have

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}_j}\Lambda_p(\Phi\boldsymbol{\beta}_n^*) &= \mathsf{E}\Big\{\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)dN(t)\Big\} \\ &\quad - \int_0^\tau \frac{\mathsf{E}[\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}{\mathsf{E}[Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]} d\,\mathsf{E}\{N(t)\} \\ &= 0. \end{aligned} \tag{69}$$

We use (69) to evaluate $S_{pj}(\boldsymbol{\beta}_n^*)$. Write

$$S_{pj}(\boldsymbol{\beta}_n^*) = J_{8j} - J_{9j} - J_{10j}, \tag{70}$$

where

$$\begin{aligned} J_{8j} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)dN_i(t) - \mathsf{E}\Big\{\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)dN(t)\Big\} \qquad (71)\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)dM_i(t)\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)Z_i(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}_i(t)\}\lambda_0(t)dt\\ &\quad - \mathsf{E}\Big[\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}\lambda_0(t)dt\Big],\\ J_{9j} &= \int_0^\tau \Big(\frac{\sum_{i=1}^n \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)Z_i(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}_i(t)\}}{\sum_{i=1}^n Z_i(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}_i(t)\}}\\ &\quad - \frac{\mathsf{E}[\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}{\mathsf{E}[Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}\Big)d\bar{N}(t), \end{aligned}$$

and

$$J_{10j} = \int_0^\tau \frac{\mathsf{E}[\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}{\mathsf{E}[Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}d[\bar{N}(t) - \mathsf{E}\{N(t)\}].$$

We consider J_{8j}, J_{9j}, and J_{10j}.
By (23), (24), and (71), we obtain

$$\begin{aligned} \mathsf{E}\{J_{8j}^2\} &\le \frac{2}{n}\mathsf{E}\Big[\int_0^\tau \{\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)\}^2 Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}\lambda_0(t)dt\Big]\\ &\quad + \frac{2}{n}\mathsf{E}\Big(\Big[\int_0^\tau \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_0^\top(t)\boldsymbol{X}(t)\}\lambda_0(t)dt\Big]^2\Big)\\ &\sim \frac{1}{n}\mathsf{E}\Big\{\int_0^\tau \boldsymbol{\phi}_j^\top(t)\Omega(t)\boldsymbol{\phi}_j(t)dt\Big\}\\ &\sim \frac{1}{n}\|\boldsymbol{\phi}_j\|_{L_2} \sim \frac{1}{n}\|\boldsymbol{\phi}_j\|^2. \end{aligned}$$

The above inequality yields

$$\mathsf{E}\Big(\sum_{j=1}^{pK_n} J_{8j}^2\Big) \le \frac{C}{n}\sum_{j=1}^{pK_n}\|\boldsymbol{\phi}_j\|^2 = \frac{CpK_n}{n}. \tag{72}$$

As in [17], we have

$$\sum_{j=1}^{pK_n} J_{9j}^2 \;\le\; \sum_{j=1}^{pK_n} \int_0^\tau \Big(\frac{\sum_{i=1}^n \boldsymbol{\phi}_j^\top(t)\boldsymbol{X}_i(t)Z_i(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}_i(t)\}}{\sum_{i=1}^n Z_i(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}_i(t)\}} - \frac{\mathsf{E}[\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}{\mathsf{E}[Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]} \Big)^2 d\bar{N}(t). \tag{73}$$

We can evaluate the expectation of (73) by following the proof of Lemma 5 of [17] with $\boldsymbol{h}_n(\boldsymbol{X}(t))$ and $\boldsymbol{\phi}_j(\boldsymbol{X}(t))$ replaced with 1 and $\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)$, respectively. Thus we have

$$\sum_{j=1}^{pK_n} \mathsf{E}\{J_{9j}^2\} \le \frac{CpK_n}{n}. \tag{74}$$

Finally we deal with J_{10j}. Since J_{10j} is just a sample mean, we have

$$\begin{aligned} &\sum_{j=1}^{pK_n} \mathsf{E}(J_{10j}^2) \qquad (75)\\ &\le \frac{1}{n}\sum_{j=1}^{pK_n} \mathsf{E}\Big\{ \Big(\int_0^\tau \frac{\mathsf{E}[\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]}{\mathsf{E}[Z(t)\exp\{\boldsymbol{g}_n^{*\top}(t)\boldsymbol{X}(t)\}]} dN(t) \Big)^2 \Big\}\\ &\le \frac{C}{n}\sum_{j=1}^{pK_n} \int_0^\tau [\mathsf{E}\{\boldsymbol{\phi}_j^\top(t)\boldsymbol{X}(t)Z(t)\}]^2 d\,\mathsf{E}\{N(t)\}\\ &\le \frac{C}{n}\sum_{j=1}^{pK_n} \int_0^\tau |\boldsymbol{\phi}_j(t)|^2 dt \;\le\; \frac{C}{n}\sum_{j=1}^{pK_n} \|\boldsymbol{\phi}_j\|_{L_2}^2 \;\le\; \frac{CpK_n}{n}. \end{aligned}$$

The desired result follows from (70), (72), (74), and (75). Hence the proof of Proposition 2 is complete.

Proof of Proposition 3. Note that $\|\Phi\boldsymbol{\beta}\|_\infty \le M$ in this proposition and take $\boldsymbol{\beta}_1 \in R^{pK_n}$. With probability tending to 1, we have uniformly in $\boldsymbol{\beta}$ and

$\boldsymbol{\beta}_1$,

$$
\begin{aligned}
&\boldsymbol{\beta}_1^\top D_p(\boldsymbol{\beta})\boldsymbol{\beta}_1 \\
&= -\int_0^\tau \Big\{ \frac{\sum_{i=1}^n Z_i(t)\boldsymbol{\beta}_1^\top \Phi^\top(t)\boldsymbol{X}_i(t)\boldsymbol{X}_i^\top(t)\Phi(t)\boldsymbol{\beta}_1 \exp[\{\Phi(t)\boldsymbol{\beta}\}^\top \boldsymbol{X}_i(t)]}{\sum_{i=1}^n Z_i(t)\exp[\{\Phi(t)\boldsymbol{\beta}\}^\top \boldsymbol{X}_i(t)]} \\
&\qquad -\Big(\frac{\sum_{i=1}^n Z_i(t)\boldsymbol{\beta}_1^\top \Phi^\top(t)\boldsymbol{X}_i(t)\exp[\{\Phi(t)\boldsymbol{\beta}\}^\top \boldsymbol{X}_i(t)]}{\sum_{i=1}^n Z_i(t)\exp[\{\Phi(t)\boldsymbol{\beta}\}^\top \boldsymbol{X}_i(t)]}\Big)^2 \Big\} d\bar{N}(t) \\
&= -\int_0^\tau \inf_c \frac{\sum_{i=1}^n Z_i(t)[\{\Phi(t)\boldsymbol{\beta}_1\}^\top \boldsymbol{X}_i(t) - c]^2 \exp[\{\Phi(t)\boldsymbol{\beta}\}^\top \boldsymbol{X}_i(t)]}{\sum_{i=1}^n Z_i(t)\exp[\{\Phi(t)\boldsymbol{\beta}\}^\top \boldsymbol{X}_i(t)]} d\bar{N}(t) \\
&\sim -\int_0^\tau \inf_c \frac{\sum_{i=1}^n Z_i(t)[\{\Phi(t)\boldsymbol{\beta}_1\}^\top \boldsymbol{X}_i(t) - c]^2}{\sum_{i=1}^n Z_i(t)} d\bar{N}(t) \\
&\sim -\|\Phi\boldsymbol{\beta}_1\|_{0n}^2 \sim -\|\Phi\boldsymbol{\beta}_1\|_0^2 \sim -\|\Phi\boldsymbol{\beta}_1\|_{L_2}^2 \sim -\|\Phi\boldsymbol{\beta}_1\|^2 \sim |\boldsymbol{\beta}_1|^2
\end{aligned}
$$

Hence the desired result is established.

Proof of Proposition 4. We have

$$(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top S_p(\boldsymbol{\beta}_1) = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top S_p(\boldsymbol{\beta}_n^*) + (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top D_p(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_n^*),$$

where $\bar{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_n^*$.

Since $\|\Phi\boldsymbol{\beta}_1\|_\infty \le M$ and $\|\Phi\boldsymbol{\beta}_n^*\|_\infty = \|\boldsymbol{g}_n^*\|_\infty \le \|\boldsymbol{g}_0\|_\infty + CA_n\rho_n$ for some positive constant C, Proposition 3 implies there are positive constants M_1 and M_2 such that with probability tending to 1,

$$-M_1 \le \lambda_{\min}(D_p(\bar{\boldsymbol{\beta}})) \le \lambda_{\max}(D_p(\bar{\boldsymbol{\beta}})) \le -M_2$$

uniformly in $\boldsymbol{\beta}_1$. Hence we obtain

$$(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^\top D_p(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_n^*) = \mathcal{O}_p(|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_n^*|)$$

uniformly in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and the proof of Proposition 4 is complete.

Proof of Lemma 5. We have uniformly in t and $\boldsymbol{h} = (h_1, \ldots, h_p)^\top \in \mathbf{H}_0$,

$$
\begin{aligned}
&\mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)\}^2 Z(t)]/\,\mathsf{E}\{Z(t)\} \\
&\quad = \mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)\}^2\,\mathsf{E}\{Z(t)|\boldsymbol{X}(t), U(t)\}]/\,\mathsf{E}\{Z(t)\} \\
&\quad \sim \mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)\}^2] \\
&\quad = \mathsf{E}\{\boldsymbol{h}^\top(U(t))\Omega(U(t), t)\boldsymbol{h}(U(t))\} \\
&\quad \sim \mathsf{E}\Big\{\sum_{j=1}^p h_j^2(U(t))\Big\} \sim \|\boldsymbol{h}\|_{L_2}^2.
\end{aligned}
$$

We used Assumptions X' and M' here. Hence

$$\begin{aligned}\|\boldsymbol{h}\|^2 &= \int_0^\tau \frac{\mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)\}^2 Z(t)]}{\mathsf{E}\{Z(t)\}} d\,\mathsf{E}\{N(t)\} \\ &\sim \|\boldsymbol{h}\|_{L_2}\,\mathsf{E}\{N(\tau)\} \sim \|\boldsymbol{h}\|_{L_2}^2\end{aligned}$$

uniformly in $\boldsymbol{h} \in \mathbf{H}_0$ and the desired result is established.

Proof of Lemma 6. First recall that the first element of $\boldsymbol{X}(t)$ is 1. We obtain as in the proof of Lemma 5,

$$\frac{\mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t) - c\}^2 Z(t)]}{\mathsf{E}\{Z(t)\}} \sim \|h_1 - c\|_{L_2}^2 + \sum_{j=2}^p \|h_j\|_{L_2}^2$$

uniformly in c, t, and $\boldsymbol{h} = (h_1, \ldots, h_p)^\top \in \mathbf{H}$. Thus we obtain

$$\begin{aligned}&\|\boldsymbol{h}\|_0^2 \\ &= \int_0^\tau \frac{1}{\mathsf{E}\{Z(t)\}} \mathsf{E}\Big(\Big[\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t) - \frac{\mathsf{E}\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)Z(t)\}}{\mathsf{E}\{Z(t)\}}\Big]^2 Z(t)\Big) \\ &\qquad \times d\,\mathsf{E}\{N(t)\} \\ &= \int_0^\tau \inf_c \Big(\frac{\mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t) - c\}^2 Z(t)]}{\mathsf{E}\{Z(t)\}}\Big) d\,\mathsf{E}\{N(t)\} \\ &\sim \int_0^\tau \inf_c \Big(\|h_1 - c\|_{L_2}^2 + \sum_{j=2}^p \|h_j\|_{L_2}^2\Big) d\,\mathsf{E}\{N(t)\} \\ &\sim \inf_c \int_0^\tau \Big(\|h_1 - c\|_{L_2}^2 + \sum_{j=2}^p \|h_j\|_{L_2}^2\Big) d\,\mathsf{E}\{N(t)\} \\ &\sim \inf_c \int_0^\tau \Big(\frac{\mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t) - c\}^2 Z(t)]}{\mathsf{E}\{Z(t)\}}\Big) d\,\mathsf{E}\{N(t)\} \\ &= \int_0^\tau \frac{\mathsf{E}[\{\boldsymbol{h}^\top(U(t))\boldsymbol{X}(t)\}^2 Z(t)]}{\mathsf{E}\{Z(t)\}} d\,\mathsf{E}\{N(t)\} \sim \|\boldsymbol{h}\|_{L_2}^2.\end{aligned}$$

We used the identifiability constraint (59) in the last line. Hence the proof of Lemma 6 is complete.

Proof of Lemma 7. We omit the details since we can prove Lemma 7 almost in the same way as Lemma 3 of [17] by just replacing $f_1(\boldsymbol{X}(t))$ and

$f_2(\boldsymbol{X}(t))$ there with $\boldsymbol{g}_1^\top(U(t))\boldsymbol{X}(t)$ and $\boldsymbol{g}_1^\top(U(t))\boldsymbol{X}(t)$, respectively. However, this does not affect the application of Lemma 10 of [16] since N_n, which is the counterpart of pK_n, also increases in [17].

Proof of Lemma 8. We omit the details since we can also prove Lemma 8 almost in the same way as Lemma 4 of [17] by just replacing $f_1(\boldsymbol{X}(t))$, $f_2(\boldsymbol{X}(t))$, and $\phi_i(\boldsymbol{X}(t))$ with $\boldsymbol{g}_1^\top(U(t))\boldsymbol{X}(t)$, $\boldsymbol{g}_2^\top(U(t))\boldsymbol{X}(t)$, and $\phi_i^\top(U(t))\boldsymbol{X}(t)$, respectively. The authors of [17] used their Lemma 5 in the proof of their Lemma 4. We can also verify their Lemma 5 by replacing $\phi_i(\boldsymbol{X}(t))$ and $h_n(\boldsymbol{X}(t))$ with $\phi_i^\top(U(t))\boldsymbol{X}(t)$ and 1, respectively.

References

[1] R. Bender, T. Augustin, M. Blettner, Generating survival times to simulate Cox proportional hazards models, Statist. in Medicine 24 (2005) 1713-1723.

[2] J. Bradic, J. Fan, J. Jiang, Regularization for Cox's proportional hazards model with NP-dimensionality, Ann. Statist. 39 (2011) 3092-3120.

[3] J. Bradic, R. Song, Gaussian oracle inequalities for structured selection in non-parametric partial likelihood, arXiv:1207.4510v1[math.ST].

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

[5] J. Cai, J. Fan, J. Jiang, H. Zhou, Partially linear hazard regression for multivariate survival data, J. Amer. Statist. Assoc. 102 (2007) 538-551.

[6] J. Cai, J. Fan, J. Jiang, H. Zhou, Partially linear hazard regression with varying coefficients for multivariate survival data, J. R. Statist. Soc. B 70 (2008) 141-158.

[7] J. Cai, J. Fan, R. Li, H. Zhou. Variable selection for multivariate failure time data, Boimetrika 92 (2005) 303-316.

[8] J. Cai, J. Fan, H. Zhou, Y. Zhou, Hazard models with varying coefficients for multivariate failure time data, Ann. Statist. 35 (2007) 324-354.

[9] Z. Cai, Y. Sun, Local linear estimation for time-dependent coefficients in Cox's regression models, Scand. J. Statist. 30 (2003) 93-111.

[10] K. Chen, H. Lin, Y. Zhou, Efficient estimation for the Cox model with varying coefficients, Biometrika 99 (2012) 379-392.

[11] P. Du, S. Ma, H. Liang, Penalized variable selection procedure for Cox models with semiparametric relative risk, Ann. Statist. 38 (2010) 2092-2117.

[12] J. Fan, H. Lin, Y. Zhou, Local partial-likelihood estimation for lifetime data, Ann. Statist. 34 (2006) 290-325.

[13] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348-1360.

[14] J. Fan, J. Lv, Nonconcave penalized likelihood with NP-dimensionality, IEEE transaction on Information Theory 57 (2011) 5467-5484.

[15] T. R. Fleming, D. P. Harrington, Counting Processes and Survival Analysis, Wiley, Hoboken, NJ, 1991.

[16] J. Z. Huang, Projection estimation in multiple regression with application to functional ANOVA models, Ann. Statist. 26 (1998) 242-272.

[17] J. Z. Huang, C. Kooperberg, C. J. Stone, Y. K. Truong, Functional ANOVA modeling for proportional hazards regression, Ann. Statist. 28 (2000) 961-999.

[18] J. Z. Huang, C. J. Stone, The L_2 rate of convergence for event history regression with time-dependent covariates, Scand. J. Statist. 25 (1998) 603-620.

[19] C. Leng, H. H. Zhang, The L_2 rate of convergence for event history regression with time-dependent covariates, J. Nonparametric Statist. 18 (2006) 417-429.

[20] H. Lian, Variable selection for high-dimensional generalized varying-coefficient models, forthcoming in Statistica Sinica, doi:10.5705/ss.2010.308.

[21] L. Meier, S. van de Geer, P. Bühlmann, The group lasso for logistic regression, J. R. Statist. Soc. B 70 (2008) 53-71.

[22] H. S. Noh, B. U. Park, Sparse variable coefficient models for longitudinal data, Statistica Sinica 20 (2010) 1183-1202.

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

[24] L. Tian, D. Zucker, L. J. Wei, On the Cox model with time-varying regression coefficients, J. Amer. Statist. Assoc. 100 (2005) 172-183.

[25] R. J. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Statist. Soc. B 58 (1996) 267-288.

[26] J. Yan, J. Hunag, Model selection for Cox models with time-varying Coefficients, Biometrics 68 (2012) 419-428.

[27] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Statist. Soc. B 68 (2006) 49-67.

[28] L. Wang, H. Li, J. Z. Huang, Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements, J. Amer. Statist. Assoc. 103 (2008) 172-183.

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

[30] F. Wei, J. Huang, H. Li Variable selection and estimation in high-dimensional varying-coefficient models, Statistica Sinica 21 (2011) 1515-1540.

[31] H. H. Zhang, G. Cheng, Y. Liu, Linear or nonlinear? Automatic structure discovery for partially linear models, J. Amer. Statist. Assoc. 106 (2011) 1099-1112.

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

[33] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418-1429.

SFB 649 Discussion Paper Series 2012

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "HMM in dynamic HAC models" by Wolfgang Karl Härdle, Ostap Okhrin and Weining Wang, January 2012.
002 "Dynamic Activity Analysis Model Based Win-Win Development Forecasting Under the Environmental Regulation in China" by Shiyi Chen and Wolfgang Karl Härdle, January 2012.
003 "A Donsker Theorem for Lévy Measures" by Richard Nickl and Markus Reiß, January 2012.
004 "Computational Statistics (Journal)" by Wolfgang Karl Härdle, Yuichi Mori and Jürgen Symanzik, January 2012.
005 "Implementing quotas in university admissions: An experimental analysis" by Sebastian Braun, Nadja Dwenger, Dorothea Kübler and Alexander Westkamp, January 2012.
006 "Quantile Regression in Risk Calibration" by Shih-Kang Chao, Wolfgang Karl Härdle and Weining Wang, January 2012.
007 "Total Work and Gender: Facts and Possible Explanations" by Michael Burda, Daniel S. Hamermesh and Philippe Weil, February 2012.
008 "Does Basel II Pillar 3 Risk Exposure Data help to Identify Risky Banks?" by Ralf Sabiwalsky, February 2012.
009 "Comparability Effects of Mandatory IFRS Adoption" by Stefano Cascino and Joachim Gassen, February 2012.
010 "Fair Value Reclassifications of Financial Assets during the Financial Crisis" by Jannis Bischof, Ulf Brüggemann and Holger Daske, February 2012.
011 "Intended and unintended consequences of mandatory IFRS adoption: A review of extant evidence and suggestions for future research" by Ulf Brüggemann, Jörg-Markus Hitz and Thorsten Sellhorn, February 2012.
012 "Confidence sets in nonparametric calibration of exponential Lévy models" by Jakob Söhl, February 2012.
013 "The Polarization of Employment in German Local Labor Markets" by Charlotte Senftleben and Hanna Wielandt, February 2012.
014 "On the Dark Side of the Market: Identifying and Analyzing Hidden Order Placements" by Nikolaus Hautsch and Ruihong Huang, February 2012.
015 "Existence and Uniqueness of Perturbation Solutions to DSGE Models" by Hong Lan and Alexander Meyer-Gohde, February 2012.
016 "Nonparametric adaptive estimation of linear functionals for low frequency observed Lévy processes" by Johanna Kappus, February 2012.
017 "Option calibration of exponential Lévy models: Implementation and empirical results" by Jakob Söhl und Mathias Trabs, February 2012.
018 "Managerial Overconfidence and Corporate Risk Management" by Tim R. Adam, Chitru S. Fernando and Evgenia Golubeva, February 2012.
019 "Why Do Firms Engage in Selective Hedging?" by Tim R. Adam, Chitru S. Fernando and Jesus M. Salas, February 2012.
020 "A Slab in the Face: Building Quality and Neighborhood Effects" by Rainer Schulz and Martin Wersing, February 2012.
021 "A Strategy Perspective on the Performance Relevance of the CFO" by Andreas Venus and Andreas Engelen, February 2012.
022 "Assessing the Anchoring of Inflation Expectations" by Till Strohsal and Lars Winkelmann, February 2012.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2012

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

023 "Hidden Liquidity: Determinants and Impact" by Gökhan Cebiroglu and Ulrich Horst, March 2012.
024 "Bye Bye, G.I. - The Impact of the U.S. Military Drawdown on Local German Labor Markets" by Jan Peter aus dem Moore and Alexandra Spitz-Oener, March 2012.
025 "Is socially responsible investing just screening? Evidence from mutual funds" by Markus Hirschberger, Ralph E. Steuer, Sebastian Utz and Maximilian Wimmer, March 2012.
026 "Explaining regional unemployment differences in Germany: a spatial panel data analysis" by Franziska Lottmann, March 2012.
027 "Forecast based Pricing of Weather Derivatives" by Wolfgang Karl Härdle, Brenda López-Cabrera and Matthias Ritter, March 2012.
028 "Does umbrella branding really work? Investigating cross-category brand loyalty" by Nadja Silberhorn and Lutz Hildebrandt, April 2012.
029 "Statistical Modelling of Temperature Risk" by Zografia Anastasiadou, and Brenda López-Cabrera, April 2012.
030 "Support Vector Machines with Evolutionary Feature Selection for Default Prediction" by Wolfgang Karl Härdle, Dedy Dwi Prastyo and Christian Hafner, April 2012.
031 "Local Adaptive Multiplicative Error Models for High-Frequency Forecasts" by Wolfgang Karl Härdle, Nikolaus Hautsch and Andrija Mihoci, April 2012.
032 "Copula Dynamics in CDOs." by Barbara Choroś-Tomczyk, Wolfgang Karl Härdle and Ludger Overbeck, May 2012.
033 "Simultaneous Statistical Inference in Dynamic Factor Models" by Thorsten Dickhaus, May 2012.
034 "Realized Copula" by Matthias R. Fengler and Ostap Okhrin, Mai 2012.
035 "Correlated Trades and Herd Behavior in the Stock Market" by Simon Jurkatis, Stephanie Kremer and Dieter Nautz, May 2012
036 "Hierarchical Archimedean Copulae: The HAC Package" by Ostap Okhrin and Alexander Ristig, May 2012.
037 "Do Japanese Stock Prices Reflect Macro Fundamentals?" by Wenjuan Chen and Anton Velinov, May 2012.
038 "The Aging Investor: Insights from Neuroeconomics" by Peter N. C. Mohr and Hauke R. Heekeren, May 2012.
039 "Volatility of price indices for heterogeneous goods" by Fabian Y.R.P. Bocart and Christian M. Hafner, May 2012.
040 "Location, location, location: Extracting location value from house prices" by Jens Kolbe, Rainer Schulz, Martin Wersing and Axel Werwatz, May 2012.
041 "Multiple point hypothesis test problems and effective numbers of tests" by Thorsten Dickhaus and Jens Stange, June 2012
042 "Generated Covariates in Nonparametric Estimation: A Short Review." by Enno Mammen, Christoph Rothe, and Melanie Schienle, June 2012.
043 "The Signal of Volatility" by Till Strohsal and Enzo Weber, June 2012.
044 "Copula-Based Dynamic Conditional Correlation Multiplicative Error Processes" by Taras Bodnar and Nikolaus Hautsch, July 2012

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

SFB 649 Discussion Paper Series 2012

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

045 "Additive Models: Extensions and Related Models." by Enno Mammen, Byeong U. Park and Melanie Schienle, July 2012.
046 "A uniform central limit theorem and efficiency for deconvolution estimators" by Jakob Söhl and Mathias Trabs, July 2012
047 "Nonparametric Kernel Density Estimation Near the Boundary" by Peter Malec and Melanie Schienle, August 2012
048 "Yield Curve Modeling and Forecasting using Semiparametric Factor Dynamics" by Wolfgang Karl Härdle and Piotr Majer, August 2012
049 "Simultaneous test procedures in terms of p-value copulae" by Thorsten Dickhaus and Jakob Gierl, August 2012
050 "Do Natural Resource Sectors Rely Less on External Finance than Manufacturing Sectors? " by Christian Hattendorff, August 2012
051 "Using transfer entropy to measure information flows between financial markets" by Thomas Dimpfl and Franziska J. Peter, August 2012
052 "Rethinking stock market integration: Globalization, valuation and convergence" by Pui Sun Tam and Pui I Tam, August 2012
053 "Financial Network Systemic Risk Contributions" by Nikolaus Hautsch, Julia Schaumburg and Melanie Schienle, August 2012
054 "Modeling Time-Varying Dependencies between Positive-Valued High-Frequency Time Series" by Nikolaus Hautsch, Ostap Okhrin and Alexander Ristig, September 2012
055 "Consumer Standards as a Strategic Device to Mitigate Ratchet Effects in Dynamic Regulation" by Raffaele Fiocco and Roland Strausz, September 2012
056 "Strategic Delegation Improves Cartel Stability" by Martijn A. Han, October 2012
057 "Short-Term Managerial Contracts and Cartels" by Martijn A. Han, October 2012
058 "Private and Public Control of Management" by Charles Angelucci and Martijn A. Han, October 2012
059 "Cartelization Through Buyer Groups" by Chris Doyle and Martijn A. Han, October 2012
060 "Modelling general dependence between commodity forward curves" by Mikhail Zolotko and Ostap Okhrin, October 2012
061 "Variable selection in Cox regression models with varying coefficients" by Toshio Honda and Wolfgang Karl Härdle, October 2012

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".